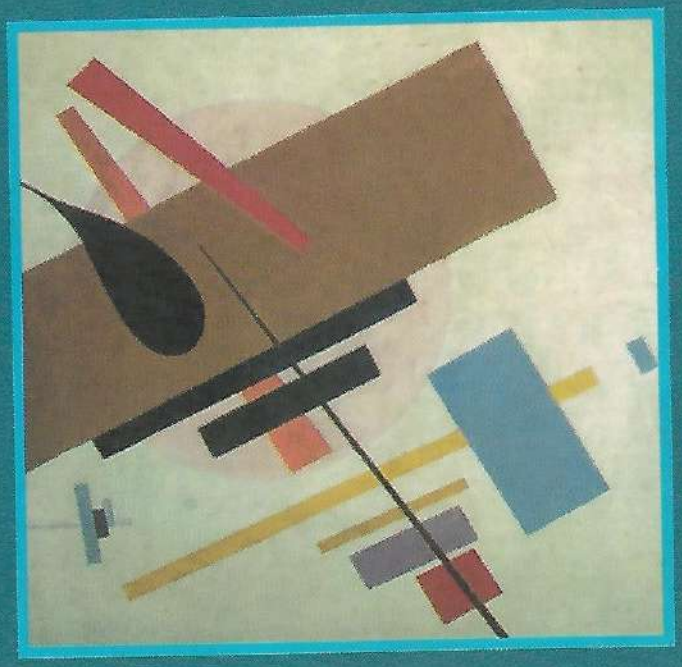
Séptima Edición

TESTS PSICOLÓGICOS



ANNE ANASTASI
SUSANA URBINA

PRENTICE HALL

Tests Psicológicos

Séptima Edición

Tests Psicológicos

ANNE ANASTASI
Department of Psychology, Fordham University

SUSANA URBINA
Department of Psychology, University of North Florida

TRADUCCIÓN:
María Elena Ortíz Salinas
Licenciada en Psicología
Universidad Nacional Autónoma de México

REVISIÓN TÉCNICA:
María Enedina Villegas Hernández
Licenciada en Psicología
Universidad Nacional Autónoma de México

PRENTICE
HALL

MÉXICO • NUEVA YORK • BOGOTÁ • LONDRES • MADRID
MUNICH • NUEVA DELHI • PARÍS • RÍO DE JANEIRO
SINGAPUR • SYDNEY • TOKIO • TORONTO • ZURICH

Datos de catalogación bibliográfica

ANASTASI ANNE, URBINA SUSANA
Tests Psicológicos
PRENTICE HALL, México, 1998

ISBN: 970-17-0186-0
Materia: Psicología

Formato: 19 x 23.5 Páginas: 744

EDICIÓN EN ESPAÑOL:

GERENTE EDITORIAL: LUIS GERARDO CEDEÑO PLASCENCIA
SUPERVISOR DE TRADUCCIÓN: JOSÉ LUIS NÚÑEZ HERREJÓN
SUPERVISOR DE EDICIÓN: JOSÉ D. HERNÁNDEZ GARDUÑO
CORRECTOR DE ESTILO: JOSÉ FRANCISCO JAVIER DÁVILA MARTÍNEZ

Edición en inglés:

Editor in Chief: Pete Janzow
Acquisitions Editor: Heidi Freund
Director of Production and Manufacturing: Barbara Kittle
Managing Editor: Bonnie Biller
Manufacturing Manager: Nick Sklitsis
Prepress and Manufacturing Buyer: Tricia Kenny
Creative Design Director: Leslie Osher
Marketing Manager: Michael Alread
Art Coordinator: Michele Giusti
Editorial Assistant: Emsal Hasan
Cover Design/Interior Design: Circa '86
Cover Photo Credit: Kasimir Malevich, *Suprematism*, c.1917, oil on canvas, 80 × 80 cm., Museum of Fine Arts, Krassnodar, Russia, Erich Lessing / Art Resources, NY
Acknowledgments begin on page 681, which constitutes a continuation of the copyright page.

ANASTASI: TESTS PSICOLÓGICOS, 7a. Ed.

Traducido de la séptima edición en inglés de la obra: **Pshychological Testing.**

All rights reserved. Authorized translation from English language edition published by Prentice-Hall, Inc. A Simon & Schuster Company.

Todos los derechos reservados. Traducción autorizada de la edición en inglés publicada por Prentice-Hall, Inc. A Simon & Schuster Company.

All rights reserved. No part of this book may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, recording or by any information storage and retrieval system, without permission in writing from the publisher.

Prohibida la reproducción total o parcial de esta obra, por cualquier medio o método sin autorización por escrito del editor.

Derechos reservados © 1998 respecto a la primera edición en español publicada por:
PRENTICE-HALL HISPANOAMERICANA, S.A.
Calle 4 No. 25-2º piso, Fracc. Industrial Alce Blanco
53370 Naucalpan de Juárez, Edo. de México
ISBN 970-17-0186-0
Miembro de la Cámara Nacional de la Industria Editorial, Reg. Núm. 1524.

Original English Language Edition Published by Prentice-Hall, Inc. A Simon & Schuster Company
Copyright © MXMXCVII
All rights reserved
ISBN 0-02-303085-2

IMPRESO EN MÉXICO/PRINTED IN MEXICO

MAY

LITOGRAFICA INGRAMEX, S.A. DE C.V.
CENTENO NO. 162-1
MEXICO, D.F.
C.P. 09810

3000 1998

Esta obra está dedicada a la memoria
de John Porter Foley, Jr., quien realizó
significativas contribuciones a cada una
de las ediciones que se han hecho sucesivamente

A.A.
S.U.

Contenido

Prefacio xi

PRIMERA PARTE

FUNCIONES Y ORÍGENES DE PRUEBAS, TESTS O INSTRUMENTOS DE MEDICIÓN PSICOLÓGICOS

1 Naturaleza y uso de las pruebas psicológicas 2

Usos y variedades de las pruebas psicológicas 2
¿Qué es una prueba psicológica? 4
¿Por qué controlar el uso de las pruebas psicológicas? 10
Aplicación de la prueba 13
Examinador y variables situacionales 17
El punto de vista del examinado 20
Efectos del entrenamiento sobre el desempeño en la prueba 23
Fuentes de información 27

2 Antecedentes históricos de las pruebas actuales 32

Interés inicial en la clasificación y la capacitación de las personas con retardo mental 33
Los primeros psicólogos experimentales 34
Contribuciones de Francis Galton 35
Cattell y los primeros tests mentales 36
Binet y el surgimiento de los tests de inteligencia 37
Pruebas colectivas 38
Tests de aptitud 39
Pruebas estandarizadas de aprovechamiento 42
Evaluación de la personalidad 44

SEGUNDA PARTE

PRINCIPIOS TÉCNICOS Y METODOLÓGICOS

3 Normas y significado de las puntuaciones de los tests 48

Conceptos estadísticos 49
Normas de desarrollo 54
Normas intragrupo 58
Relatividad de las normas 66
Las computadoras y la interpretación de las calificaciones de las pruebas 74
Interpretación de los tests referidos a dominio 76
Calificaciones mínimas y puntuaciones de corte 80

4 Confiabilidad 84

El coeficiente de correlación 85
Tipos de confiabilidad 91
Confiabilidad de las pruebas de velocidad 102
Dependencia de los coeficientes de confiabilidad de la muestra examinada 105
Error estándar de medición 107
Aplicación de la confiabilidad a las pruebas de destreza y las puntuaciones de corte 112

5 Validez: conceptos básicos 113

Evolución de los conceptos de validez de las pruebas 114
Procedimientos de la descripción del contenido 114
Procedimientos de criterio-predicción 118
Procedimientos de identificación del constructo 126
Recaptitulación e integración 136

6 Validez: medición e interpretación 140

Coeficiente de validez y error de estimación 141
Validez de test y teoría de la decisión 144
Combinación de información a partir de diferentes tests 156
Uso de los tests para decisiones de clasificación 160
Análisis estadístico del sesgo de la prueba 164

7 Análisis de reactivos 172

Dificultad de los reactivos 173
Discriminación del reactivo 179
Teoría de respuesta al ítem 187
Análisis de reactivos de las pruebas de velocidad 193
Validación cruzada 194
Funcionamiento diferencial de los reactivos 196
Exploraciones en el desarrollo de reactivos 200

TERCERA PARTE

TESTS DE HABILIDAD

8 Pruebas individuales 204

Escala de inteligencia Stanford-binet 205
Las escalas de Wechsler 214
Las escalas de Kaufman 222
Escalas de habilidad diferencial 226
Sistema de evaluación cognoscitiva de Das-Naglieri 233

9 Pruebas para poblaciones especiales 234

Pruebas de infantes y preescolares 235

Evaluación extensa de personas con retardo mental 247
Examinación de personas con discapacidades físicas 252
Exámenes multiculturales 259

10 Pruebas colectivas 271

Las pruebas colectivas en comparación con las individuales 272
Evaluación adaptativa y aplicación computarizada 274
Baterías de niveles múltiples 278
Medición de aptitudes múltiples 287

11 Naturaleza de la inteligencia 294

Significado del CI 295
Heredabilidad y modificabilidad 297
Motivación e inteligencia 300
Análisis factorial de la inteligencia 303
Teorías de la organización de los rasgos 309
Naturaleza y desarrollo de los rasgos 318

12 Consecuencias psicológicas en la evaluación de la habilidad 323

Estudios longitudinales de la inteligencia infantil 323
La inteligencia en la niñez temprana 327
Problemas en la evaluación de la inteligencia de los adultos 331
Cambios poblacionales en la ejecución en los tests de inteligencia 337
Diversidad cultural 340

CUARTA PARTE

EVALUACIÓN DE LA PERSONALIDAD

13 Inventarios autodescriptivos de personalidad 348

Procedimientos relacionados con el contenido 349
Clave empírica de criterio 350
El análisis factorial en el desarrollo de pruebas 362
La teoría de la personalidad en el desarrollo de las pruebas 367
Las actitudes del examinado y los sesgos de respuesta 374
Rasgos, estados, personas y situaciones 379
Estado actual de los invetarios de personalidad 385

14 Medición de intereses y actitudes 386

Inventario de intereses: situación actual 387
El invetario de intereses de Strong 389
Inventarios de intereses: recapitulación y algunos puntos destacados 396
Algunas tendencias significativas 402
Encuestas de opinión y escalas de actitud 404
Locus de control 408

15 Técnicas proyectivas 410

Naturaleza de las técnicas proyectivas 411
Técnicas de las manchas de tinta 411
Técnicas pictóricas o gráficas 419
Técnicas verbales 425
Recuerdos autobiográficos 427
Técnicas de ejecución 429
Evaluación de las técnicas proyectivas 432

16 Otras técnicas de evaluación 443

Medidas de estilos y tipos 443
Pruebas situacionales 450
Autoconcepto y constructos personales 454
Informes de los observadores 463
Datos biográficos 469

QUINTA PARTE

APLICACIONES DE LAS PRUEBAS

17 Principales contextos del uso actual de las pruebas 474

Evaluación educativa 474
Evaluación ocupacional 491
Uso de las pruebas en la psicología clínica y la consejería 510

18 Consideraciones éticas y sociales de la evaluación 533

Aspectos éticos en la evaluación y examinación psicológica 537
Capacidades y competencia profesional del usuario 538
Responsabilidades de los editores de las pruebas 540
Protección de la privacidad 542
Confidencialidad 544
Comunicación de los resultados de las pruebas 545
Evaluación en diversas poblaciones 546

APÉNDICES

A. Lista por orden alfabético de pruebas y otros instrumentos de evaluación 553

B. Direcciones de editoriales, distribuidores y organizaciones relacionados con las pruebas 557

Bibliografía 563
Reconocimientos 682
Índice de autores 685
Índice temático 689
Índice de instrumentos 726

Prefacio

La década de los noventa ha presenciado la continuación y el crecimiento del renovado interés en las pruebas psicológicas que se hizo evidente en los años ochenta, como lo muestra el desarrollo de recientes pruebas —algunas de las cuales representan nuevos planteamientos— así como la revisión y la continuada investigación de las anteriores. Al elegir los instrumentos que queríamos citar o analizar, la primera meta era exponer al lector a la rica variedad de herramientas de medición de que se dispone actualmente en el campo, lo mismo que a algunas pruebas y técnicas de importancia histórica; cualquier tentativa por cubrir exhaustivamente el campo, o incluso una parte significativa, es una tarea que desde luego rebasa el alcance de este libro.

Cada vez es mayor la atención que se concede al sujeto examinado, por lo que se exhorta al usuario de las pruebas a investigar, en el caudal de experiencias y reacciones propias, las causas de su desempeño: ¿Qué hay en los antecedentes del individuo que nos ayude a entender sus respuestas a la prueba?, ¿cuáles de sus elementos pueden aumentar el valor de los resultados para predecir su conducta ulterior en la escuela, el trabajo y otras actividades cotidianas? De lo anterior se desprende que ha aumentado la responsabilidad del *usuario de la prueba* en cuanto a la selección de los instrumentos adecuados, los métodos para presentarlos al individuo, la interpretación de los resultados y la comunicación y utilización de los mismos. Debido a dichas consideraciones, la obra está especialmente diseñada para brindar una base que permita el uso adecuado de las pruebas.

El empleo eficaz de los instrumentos requiere de cierta familiaridad con su elaboración, pues esta información es necesaria para evaluar distintas pruebas, elegir la apro-

piada para determinado propósito o sujeto e interpretar adecuadamente las puntuaciones. Aunque este libro no se dirige en particular a quienes elaboran los instrumentos, incluye información suficiente sobre su realización para satisfacer las necesidades del usuario.

Esta edición también proporciona explicaciones sencillas de algunos conceptos y procedimientos de uso común y rápido desarrollo que tal vez influyan en las prácticas psicométricas del siglo XXI, como es el caso de los tests adaptados a las computadoras (*computerized adaptive testing*, CAT), el metanálisis, el modelamiento de ecuaciones estructurales, el uso de intervalos de confianza en lugar de la significancia estadística tradicional, las pruebas transculturales y el uso creciente del análisis factorial en el desarrollo de los tests de habilidad y de personalidad. Una aplicación práctica del método de análisis de factores ha sido el establecimiento de normas que permitan la interpretación de las puntuaciones en diferentes niveles de especificidad o generalidad, de modo que el usuario aplique el más adecuado a la persona o situación particular.

En el uso actual de las pruebas se han hecho muy evidentes dos tendencias de gran importancia a largo plazo; en lugar de dedicarles capítulos aparte, decidimos analizarlas siempre que resultó conveniente en todo el libro. La primera es la creciente influencia de la computación en el desarrollo, la elaboración y aplicación de los instrumentos, además de los usos conocidos de las computadoras en la calificación de las pruebas y el procesamiento de las puntuaciones. La velocidad de los avances tecnológicos es tan sorprendente que parece marcar el ritmo del progreso en áreas importantes de la psicología. Asimismo, la tecnología también hace contribuciones notables a los frentes teóricos y de investigación; por ejemplo, en la bibliografía psicológica se observa una gran cantidad de integración y fecundación de distintas áreas, situación que se ve favorecida por la facilidad con la que los investigadores de todo el mundo tienen acceso a la información, la procesan y la comunican. Uno de los casos más destacados y prometedores de esta tendencia a la integración es el replanteamiento de los rasgos cognoscitivos y de personalidad como aspectos inseparables del individuo, que a su vez está inextricablemente ligado a su yo físico, a los sucesos de su vida y a su ambiente.

La segunda tendencia significativa en las pruebas psicológicas es la creciente intrusión de intereses políticos y legales. Aunque se trata de una influencia que provoca divisiones y, por ende, es potencialmente nociva, también puede tener ramificaciones positivas pues fomenta la creatividad y aumenta la vigilancia de las consecuencias, deliberadas o no, del uso de las pruebas. A lo largo del texto se citan algunas leyes promulgadas en los Estados Unidos que han tenido un gran efecto en las prácticas psicométricas, incluidos el título y el año en que fueron aprobadas; su contenido puede encontrarse en el *Congressional Record* de los Estados Unidos, así como en otras publicaciones periódicas en las secciones de referencia de la mayor parte de las bibliotecas de ese país.

Las primeras seis ediciones de este libro fueron escritas por una sola persona; en cambio, ésta es producto de una verdadera colaboración. Las dos autoras planearon la reorganización de los capítulos y los temas principales. La revisión y reelaboración de los capítulos se distribuyó de la siguiente manera: Anastasi fue responsable de los capítulos 1 a 7 y 10 a 12, mientras que Urbina se encargó de los capítulos 8, 9 y del 13 al 18 y asumió mayores funciones administrativas y de correspondencia. Además, cada

una leyó el borrador de los capítulos de la otra y sugirió mejoras que en general fueron aceptadas.

Obviamente, este libro no se hubiera escrito sin el acceso a las investigaciones y las publicaciones de muchos psicólogos de diversas partes del mundo, tanto contemporáneos como de épocas pasadas. Sus nombres aparecen en las citas de sus publicaciones y en la lista bibliográfica que se encuentra al final de la obra. Dentro de este grupo impresionante destacan algunos individuos por su puntual cooperación y la magnitud de sus contribuciones. Entre ellos sobresalen Dianne Brown del Directorio Científico de la Asociación Psicológica Estadounidense (*American Psychological Association*, APA), Aurelio Prifitera y Joanne Lenke de la *Psychological Corporation*, Lorin Letendre de *Consulting Psychologists Press*, Carol Watson de *National Computer Systems*, Douglas Jackson de *Sigma Assessment Systems*, Elizabeth McGrath y John Oswald de la *Riverside Publishing Company* y Wayne Camara del *College Board*. Además, reconocemos con gratitud la ayuda que nos prestaron los miembros del personal de las bibliotecas de las universidades Fordham y del Norte de Florida para enfrentar los retos que presentó la preparación de esta obra.

A.A.
S.U.

PRIMERA PARTE

Funciones y orígenes de pruebas, tests o instrumentos de medición psicológicos

Capítulo Uno

Naturaleza y uso de las pruebas psicológicas

Las pruebas, tests o instrumentos de medición psicológicos son herramientas, y para obtener los beneficios que proporcionan es necesario tener presente este hecho esencial. Cualquier herramienta puede ser un medio para hacer el bien o el mal, dependiendo de cómo se emplee. Las pruebas se han desarrollado a un paso creciente, y aunque cada vez son más las áreas de la vida cotidiana a las que contribuyen,[1] este crecimiento ha estado acompañado de algunos abusos y de expectativas poco realistas. El usuario de los tests necesita saber cómo evaluarlos. ¿Qué tan buena es esta prueba para el propósito que se pretende que cumpla? ¿Qué información puede brindar sobre la persona a la que se aplica? ¿Cómo pueden integrarse sus resultados en la red de datos que se utiliza en la toma de decisiones? Escribimos este libro desde el punto de vista de estas preguntas, por lo que no se dirige al especialista, sino más bien al estudiante de psicología. En la actualidad se requiere de ciertos conocimientos básicos sobre los instrumentos de medición no sólo entre quienes los elaboran o aplican, sino también de parte de cualquiera que se sirva de sus resultados como fuente de datos para tomar decisiones acerca de sí mismo o de los demás.

USOS Y VARIEDADES DE LAS PRUEBAS PSICOLÓGICAS

Habitualmente, la función de las pruebas psicológicas ha sido medir las diferencias entre individuos o entre las reacciones de la misma persona en circunstancias distintas. El diagnóstico del retardo mental fue uno de los primeros problemas que esti-

[1]Véase Dahlstrom (1993b) para una lúcida ilustración de las contribuciones de las pruebas psicológicas con ejemplos reales.

mularon su desarrollo, y, hasta el momento, la detección de las deficiencias intelectuales sigue siendo una aplicación importante de ciertos instrumentos. Los usos clínicos incluyen el examen de personas con trastornos emocionales graves y otros problemas de conducta. La evaluación de las necesidades educativas dio un fuerte impulso al desarrollo inicial de las pruebas, como fue el caso de los famosos tests de Binet que comenzaron el movimiento de las mediciones de la inteligencia. En la actualidad, las escuelas se cuentan entre los principales usuarios, ya que les permiten, entre otras muchas cosas, clasificar a los niños según su capacidad para beneficiarse de las diferentes formas de instrucción escolar, identificar a los excepcionalmente lentos o rápidos para aprender, brindar asesoría educativa y vocacional a los estudiantes de educación media y superior, y seleccionar a los aspirantes a las escuelas profesionales.

La selección y clasificación del personal industrial es otra aplicación fundamental de las pruebas psicológicas. Desde el operador de la línea de montaje y el archivista, hasta las funciones directivas, difícilmente puede encontrarse un puesto para el que alguna prueba no haya demostrado su utilidad psicológica en cuanto a contratación, asignación de tareas, transferencias, ascensos o despidos. En muchas de estas situaciones, en especial cuando se relacionan con los puestos de nivel superior, se requiere que las pruebas se empleen junto con una entrevista realizada por un experto que, al interpretar las puntuaciones a la luz de otra información importante sobre el individuo, las aprovecha mejor. Con todo, la aplicación de pruebas constituye una parte importante del programa global del departamento de personal. Una aplicación relacionada de las pruebas psicológicas se encuentra en la selección y clasificación del personal militar. Luego de sus inicios en la Primera Guerra Mundial, el alcance y la variedad de los instrumentos de medición psicológicos usados en contextos militares mostraron un desarrollo notable durante la Segunda Guerra Mundial. Posteriormente, su investigación y desarrollo ha continuado a gran escala y en todas las ramas de las fuerzas armadas.

En la consejería individual el uso de pruebas ha aumentado gradualmente de una orientación limitada a los planes educativos y vocacionales al interés en todos los aspectos de la vida de la persona. El bienestar emocional y las relaciones personales adecuadas se han convertido en objetivos prominentes de la consejería. También se observa una tendencia a servirse de las pruebas para aumentar el desarrollo y la comprensión personales. En este marco, las puntuaciones de los instrumentos son parte de la información que se proporciona al individuo para ayudarlo a tomar decisiones.

Resulta evidente el uso de los instrumentos de medición psicológicos en la solución de una gran variedad de problemas prácticos; sin embargo, no hay que perder de vista el hecho de que las pruebas también cumplen funciones importantes en la investigación básica. Por ejemplo, en casi todos los problemas de psicología diferencial se aplican tests para obtener datos, como ocurre con los estudios sobre la naturaleza y el grado de las diferencias individuales, la organización de los rasgos psicológicos, la medición de las diferencias grupales y la identificación de los factores biológicos y culturales asociados con las variaciones conductuales. En todas estas áreas de investigación —como en muchas otras— es fundamental la medición precisa de las diferencias individuales que las pruebas bien formuladas hacen posible. Del mismo modo, las pruebas

psicológicas proporcionan herramientas estandarizadas para investigar problemas tan diversos como los cambios que sufre el individuo a lo largo del ciclo de desarrollo, la eficacia relativa de distintos procedimientos educativos, los resultados de la psicoterapia, el impacto de los programas comunitarios y la influencia de las variables ambientales en el desempeño.

Las pruebas diseñadas para estos diversos propósitos también difieren en otras características notables. Varían en la forma en que se aplican, ya sea que el examinador capacitado trabaje con un individuo y luego con otro o de manera simultánea con grupos grandes, o bien por medio de una computadora. También difieren en los aspectos de la conducta que cubren. Algunas se concentran en la evaluación de los rasgos cognoscitivos o las habilidades, que pueden ir de aptitudes generales —como la capacidad de beneficiarse de la educación universitaria— a las habilidades sensoriomotoras muy especializadas que se requieren para realizar una operación manual sencilla. Otros instrumentos miden las variables afectivas o de personalidad, como los rasgos emocionales o motivacionales, la conducta interpersonal, los intereses, las aptitudes y los valores.

Frente a tal diversidad de naturaleza y propósito, ¿qué características tienen en común las pruebas psicológicas?, ¿en qué se diferencian de otros métodos para obtener información acerca de los individuos? La respuesta se encuentra en ciertos rasgos fundamentales de su elaboración y uso, que constituyen el punto de interés de este capítulo.

¿QUÉ ES UNA PRUEBA PSICOLÓGICA?

Una muestra de conducta. En esencia, la prueba psicológica es una medida objetiva y estandarizada de una muestra de conducta. Con las pruebas psicológicas, como con las de cualquier otra ciencia, se hacen observaciones sobre una *muestra* pequeña, pero cuidadosamente elegida, de la conducta del individuo. A este respecto, el psicólogo procede de la misma manera que el bioquímico que analiza la sangre de un paciente o el suministro de agua de la comunidad, examinando una o más muestras. Si el psicólogo desea probar el léxico de un niño, la habilidad de un escolar para realizar cálculos aritméticos o la coordinación visomotora de un piloto, prueba su desempeño en un conjunto representativo de palabras, problemas aritméticos o pruebas motrices. Que el instrumento cubra adecuadamente o no la conducta considerada depende obviamente del número y la naturaleza de los reactivos de la muestra; por ejemplo, una prueba de aritmética que no tenga más que cinco problemas o que sólo incluya multiplicaciones sería una mala medida de la habilidad del sujeto para hacer operaciones. Una prueba de vocabulario compuesta en exclusiva por términos provenientes de la jerga del béisbol difícilmente podría brindar una estimación confiable del léxico del niño.

El *valor de diagnóstico o predictivo* de un test psicológico depende de qué tanto funcione como indicador de un área de conducta relativamente amplia y significativa. La medición de la muestra de conducta que examina la prueba rara vez, si acaso, es el objetivo. El conocimiento del niño de una lista particular de 50 palabras no es, en sí, de gran interés, como tampoco es de mayor importancia el desempeño de quien soli-

cita un empleo en un conjunto de 20 problemas aritméticos. Pero los tests cumplen su propósito si muestran que hay una correspondencia estrecha entre el conocimiento que el niño tiene de la lista de palabras y su dominio del vocabulario o entre la puntuación que obtiene el solicitante en los problemas aritméticos y su desempeño en el empleo.

A este respecto, hay que observar que no es necesario que los reactivos se asemejen a la conducta que la prueba pretende predecir, lo único que se requiere es demostrar una correspondencia empírica entre ambos, de ahí que el grado de similitud entre la muestra de la prueba y la conducta por predecir varíe mucho. En un extremo, la prueba puede coincidir por completo con una parte de la conducta que se quiere predecir, como en el caso de una prueba de vocabulario de una lengua extranjera que examina al estudiante en 20 de las 50 palabras estudiadas o el de la prueba de conducción para obtener la licencia de manejo. Un grado menor de semejanza se encuentra en muchas pruebas de aptitud vocacional que se aplican antes de la capacitación para el trabajo, en las que hay apenas un parecido moderado entre las tareas que se realizan en el puesto y las que incluye la prueba. En el otro extremo se encuentran los tests proyectivos de personalidad, como el de manchas de tinta de Rorschach, en el que a partir de las asociaciones que el examinado hace de las manchas se intenta predecir su reacción a otras personas, a estímulos emocionales y a otras situaciones complejas de la vida cotidiana. A pesar de sus diferencias superficiales, todas esas pruebas constan de muestras de conducta del individuo, y cada una debe probar su valor con la demostración de una correspondencia empírica entre el desempeño del examinado en la prueba y en otras situaciones.

También representa una distinción menor que se utilice el término "diagnóstico" o "predicción". Por lo general, el segundo connota una estimación temporal; por ejemplo, se pronostica el desempeño del individuo en un trabajo a partir de su ejecución en la prueba. Pero en un sentido más amplio incluso el diagnóstico de una condición actual, como el retardo mental o un trastorno emocional, lleva implícita la predicción de lo que el individuo hará en otras situaciones. Lógicamente, es más sencillo considerar todas las pruebas como muestras de conducta a partir de las cuales se hacen predicciones que atañen a otro comportamiento. Entonces, es posible caracterizar las diferentes clases de pruebas o tests como variaciones de este patrón básico.

Otro punto que debemos considerar desde el principio tiene que ver con el concepto de *capacidad*. Por ejemplo, es totalmente posible elaborar una prueba para predecir qué tan bien aprenderá el francés un alumno antes de que empiece el curso. La prueba debería incluir una muestra de los comportamientos que se requieren para aprender el nuevo idioma y también presuponer que el estudiante no tiene ningún conocimiento. En este caso podría decirse que la prueba mide la "capacidad" o "potencialidad" del individuo para aprender el francés. Sin embargo, hay que tener cautela al emplear esos términos en relación con las pruebas psicológicas. Únicamente podemos decir que una prueba mide la "capacidad" en el sentido de que una muestra de la conducta actual puede utilizarse como indicador de otra conducta futura. Ninguna prueba psicológica puede hacer más que medir el comportamiento, y que éste sirva como índice de otra conducta sólo lo establece un experimento empírico.

Estandarización. Recordemos que, en la definición inicial, dijimos que la prueba psicológica es una medida estandarizada. La estandarización supone la *uniformidad de los procedimientos* en la aplicación y calificación de la prueba. Es evidente que si los resultados que obtienen distintas personas han de ser comparables, las condiciones del examen tienen que ser las mismas para todos. Tal requisito es sólo una manifestación de la necesidad de tener condiciones controladas en todas las observaciones científicas. En una situación de prueba, la única variable independiente es a menudo el individuo examinado.

Para asegurar la uniformidad de las condiciones de prueba, quien la elabora proporciona instrucciones detalladas para la aplicación de cada nuevo instrumento. La formulación de las instrucciones es una parte importante de la estandarización de la nueva prueba, y se extiende a los materiales exactos que debe emplearse, los límites de tiempo, las instrucciones orales, las demostraciones previas, las formas de manejar las dudas de los examinados y cualquier otro detalle de la situación de examinación. Así, al dar instrucciones o presentar oralmente los problemas, hay que considerar la velocidad con que se habla, el tono de la voz, la inflexión, las pausas y la expresión del rostro; por ejemplo, en una prueba que consiste en detectar absurdos, uno puede dar la respuesta correcta al sonreír o al hacer una pausa cuando se lee la palabra crucial. En una sección posterior del capítulo, que trata de los problemas de la aplicación de las pruebas, veremos el procedimiento de estandarización.

Otro paso importante en la estandarización de las pruebas es el establecimiento de *normas*. Las pruebas psicológicas no tienen criterios predeterminados de aprobación o reprobación; el desempeño en cada prueba se evalúa sobre la base de los datos empíricos. Para la mayor parte de los propósitos, a fin de interpretar el resultado que obtiene el individuo en una prueba, éste se compara con los resultados de otros en la misma prueba. Como lo sugiere el término, la norma es el desempeño normal o promedio. De esta manera, si en una prueba de razonamiento aritmético los niños normales de ocho años resuelven correctamente 12 de 50 problemas, entonces, en esta prueba, la norma para los ocho años corresponde a una puntuación de 12. Esto se conoce como puntuación cruda (bruta o directa), y se expresa como el número de reactivos correctos, el tiempo requerido para completar la tarea, el número de errores o alguna otra medida objetiva que sea adecuada para el contenido de la prueba. Esta puntuación cruda no tiene sentido hasta que no se evalúa en términos de datos interpretativos adecuados.

Durante el proceso de estandarización, la prueba se aplica a una muestra grande y representativa de las personas a las que va dirigida. Este grupo, conocido como muestra de estandarización, sirve para establecer las normas, que indican no sólo el desempeño promedio sino también la frecuencia relativa de las desviaciones por encima y por debajo del promedio, lo que permite evaluar diferentes grados de superioridad e inferioridad. En el capítulo 3 consideraremos las formas concretas de expresar tales normas, que permiten designar la posición del individuo en relación con la muestra normativa o de estandarización.

También conviene observar que, para los tests de personalidad, las normas se establecen esencialmente de la misma manera que para los de aptitud. En uno de personalidad la norma no es por fuerza la ejecución más deseable o "ideal", como tampoco

una puntuación perfecta o sin errores es la norma de un test de aptitud. En ambas pruebas la norma corresponde a la ejecución de la persona promedio. Por ejemplo, en las pruebas de dominancia-sumisión, la norma cae en un punto medio que representa el grado de dominio o de sumisión que manifiesta la persona promedio. De modo similar, en un inventario de ajuste emocional la norma por lo general no corresponde a una ausencia absoluta de respuestas inadaptadas o desfavorables. La mayoría de los individuos "normales" de la muestra de estandarización presenta algunas de esas respuestas, y este número de respuestas inadaptadas, por consecuencia, podría representar la norma.

Medición objetiva de la dificultad. Al iniciar este análisis definimos las pruebas psicológicas como una medición objetiva y estandarizada. ¿En qué sentido específico decimos que es objetiva? Ya tocamos algunos aspectos de la objetividad al hablar de la estandarización. En efecto, la aplicación, calificación e interpretación de los resultados serán objetivas en la medida en que sean independientes del juicio subjetivo del examinador. En teoría, cualquier individuo al que se aplique la prueba puede obtener una puntuación idéntica independientemente de quién la aplique. Por supuesto, esto no es del todo cierto porque en la práctica no se han alcanzado la estandarización ni la objetividad perfectas. Pero al menos la objetividad es la meta de la elaboración de instrumentos y casi todos la demuestran en un grado razonablemente elevado.

Hay otras condiciones que permiten señalar a las pruebas psicológicas como objetivas. La determinación del grado de dificultad de un reactivo o de toda la prueba se basa en procedimientos objetivos empíricos. Cuando Binet y Simon prepararon en 1905 su escala original para la medición de la inteligencia, distribuyeron los 30 reactivos de la escala en orden de dificultad creciente, que determinaron luego de probar los reactivos en 50 niños normales y en algunos con retardo mental. Tomaron los reactivos que resolvió correctamente el mayor número de niños, *ipso facto*, como los más sencillos, mientras que consideraron más difíciles los que pocos solucionaron. Con este procedimiento, establecieron un orden empírico de dificultad. Este primer ejemplo es característico de la medición objetiva del nivel de dificultad, que ahora es una práctica común en la elaboración de pruebas psicológicas.

No sólo el ordenamiento, sino también la selección de reactivos para su inclusión en una prueba, puede apoyarse en la proporción de sujetos de la muestra que resuelve cada reactivo. Así, si hay muchos reactivos en el extremo sencillo o el difícil de la escala, es posible descartar algunos. De modo similar, si en ciertas partes de la graduación de la dificultad de los reactivos son escasos, es posible agregar otros para llenar las lagunas. En el capítulo 7 trataremos aspectos más técnicos del análisis de reactivos.

Confiabilidad. ¿Qué tan buena es la prueba? ¿En realidad funciona? Estas preguntas podrían —y ocasionalmente lo hacen— exigir largas horas de análisis infructuoso. Por una parte, las opiniones subjetivas, las corazonadas y los sesgos personales pueden conducir a afirmaciones extravagantes respecto de lo que puede lograr una prueba o, por la otra, a un rechazo obstinado. La única forma de que estas preguntas reciban una respuesta concluyente es realizar una comprobación empírica. La *evaluación objetiva*

de las pruebas psicológicas consiste principalmente en determinar su confiabilidad y validez en situaciones especificadas.

Como se emplea en la psicometría, el término "confiabilidad" significa básicamente consistencia. La confiabilidad de una prueba es la consistencia de las puntuaciones obtenidas por las mismas personas cuando se les aplica la misma prueba o una forma equivalente. Si un niño tiene un CI de 110 el lunes y uno de 80 el viernes, es obvio que no se puede confiar mucho en ninguna de las dos puntuaciones. Asimismo, si, de un grupo de 50 palabras, el examinado identifica bien 40, mientras que, de otro grupo supuestamente equivalente, obtiene una puntuación de 20 correctas, ninguna de las puntuaciones puede considerarse como un indicador confiable de su comprensión verbal. Es posible que en ambos ejemplos sólo una de las puntuaciones sea errónea, pero esto sólo se demuestra con exámenes posteriores. De los datos obtenidos, lo único que se concluye es que no pueden ser correctos ambos, y sin información adicional es imposible establecer que uno o ninguno sea una estimación adecuada de la habilidad del individuo.

Antes de permitir la libre circulación de una prueba psicológica debe llevarse a cabo una verificación cuidadosa y objetiva de su confiabilidad. En el capítulo 4 estudiaremos los distintos tipos de confiabilidad, así como los métodos de medición de cada uno. Para comprobar la confiabilidad se comparan las puntuaciones obtenidas por las mismas personas en diversos momentos con diferentes conjuntos de reactivos, examinadores o calificadores, o en cualquier otra condición de examinación pertinente. Es esencial especificar el tipo de confiabilidad y el método empleado para determinarlo, ya que la misma prueba puede variar en esos diferentes aspectos. También hay que informar del número y la clase de personas con las que se hizo la verificación. Con estos datos, los usuarios pueden predecir si la prueba será tan confiable para el grupo al que esperan aplicarla, o si es probable que sea mayor o menor.

Validez. Es indudable que la pregunta más importante sobre cualquier prueba psicológica atañe a su validez —es decir, el grado con el que verdaderamente mide lo que pretende medir—. La validez proporciona una comprobación directa de qué tan bien cumple una prueba su función. Por lo general, para determinarla se requiere de *criterios* independientes y externos de lo que la prueba intenta medir. Por ejemplo, si se quiere emplear una prueba de aptitud médica para seleccionar, entre los aspirantes para ingresar a la escuela de medicina, a los más prometedores, un criterio puede ser el éxito de los seleccionados en la escuela. Durante el proceso de validación, la prueba debe aplicarse a un grupo grande de estudiantes en el momento de su admisión. Posteriormente tiene que obtenerse una medida del desempeño académico de cada uno sobre la base de sus notas, la calificación que reciban de los profesores, la terminación o el abandono de sus estudios y cosas similares. Esta medida constituye el criterio con el que se correlaciona la puntuación que recibió al principio cada estudiante. Una correlación, o *coeficiente de validez*, elevada significa que los individuos que en la prueba obtuvieron una calificación relativamente alta han sido más o menos exitosos en la escuela de medicina, mientras que los que obtuvieron bajas calificaciones en la prueba han tenido un pobre desempeño académico. Una correlación baja indicaría que existe poca correspondencia entre la puntuación

obtenida en la prueba y la medida considerada como criterio y, por lo tanto, que la validez de la prueba es poca. El coeficiente de validez nos permite determinar qué tan bien se predice el desempeño que se toma como criterio a partir de las puntuaciones de la prueba.

Las pruebas diseñadas para otros propósitos se validan de manera similar contra criterios apropiados. Por ejemplo, una prueba de aptitud vocacional puede validarse con el éxito laboral de un grupo experimental de nuevos empleados; una batería de aptitud para pilotos, con los resultados en los vuelos de entrenamiento. Las pruebas destinadas a usos más amplios y variados se validan con una serie de indicadores conductuales obtenidos de modo independiente, y su validez sólo puede establecerse con la acumulación gradual de datos de muchas investigaciones diferentes

Tal vez el lector haya notado una paradoja aparente en el concepto de validez de la prueba. Si es necesario hacer un seguimiento de los examinados u obtener de otro modo medidas independientes de lo que la prueba pretende medir, ¿por qué no prescindir de la prueba? La respuesta a este acertijo se encuentra en la distinción entre el grupo de validación por un lado y, por otro, los grupos a los que se aplicará la prueba con propósitos operativos. Antes de que la prueba esté lista para su uso es necesario establecer su validez con una muestra representativa de personas cuyas calificaciones no se emplean con propósitos operativos, sino sólo en el proceso de comprobación del instrumento. Si la prueba demuestra ser válida con ese método, puede utilizarse con otras muestras en ausencia de las medidas de criterio.

Aún podría argumentarse que sólo se necesita esperar a que la medida de criterio madure —que esté disponible— en *cualquier* grupo para obtener la información que la prueba trata de predecir. Pero semejante procedimiento supone un desperdicio tal de tiempo y energía que resultaría prohibitivo en casi todos los casos. Así, para determinar qué solicitantes tienen éxito en un empleo o qué estudiantes terminan con éxito la universidad, admitiríamos a todo aquel que lo solicite (o a una muestra aleatoria) y esperaríamos a ver que pasé. Las pruebas están diseñadas para disminuir al mínimo el derroche que supone este procedimiento, así como su nocivo impacto emocional en los individuos. Por medio de las pruebas es posible evaluar, con un margen de error determinable, el nivel actual de la persona en las habilidades requeridas, sus conocimientos así como otras características pertinentes. Entre mayor sea la validez y la confiabilidad de la prueba, menor será el margen de error.

En los capítulos 5 y 6, estudiaremos tanto los problemas especiales que uno enfrenta al determinar la validez de diversas pruebas como los criterios y los procedimientos estadísticos utilizados; sin embargo, en este momento es necesario considerar otro punto. La validez no sólo nos indica el grado en que la prueba cumple con su función, pues al estudiar los datos de la validación podemos determinar con objetividad *qué* es lo que mide el instrumento. En consecuencia, sería más preciso definir la validez como el grado en que sabemos qué es lo que mide la prueba. La interpretación de las puntuaciones sería indudablemente más clara y menos ambigua si las pruebas recibieran su nombre de acuerdo con las relaciones empíricas que las validaron. Se observa una tendencia en esta dirección en nombres como "prueba de evaluación académica" y "prueba de clasificación de personal" en lugar de títulos tan vagos como "test de inteligencia".

¿POR QUÉ CONTROLAR EL USO DE LAS PRUEBAS PSICOLÓGICAS?

"¿Puede venderme un ejemplar del Stanford-Binet? La próxima semana mi sobrino debe presentar una de las pruebas para ser admitido en la escuela X y necesita practicar para poder pasarla."

"Para mejorar el programa escolar de lectura necesitamos una prueba de CI justa que mida el potencial innato de cada niño."

"Anoche contesté las preguntas de un test de inteligencia publicado en una revista y obtuve un CI de 80. Me parece que las pruebas psicológicas no tienen sentido."

"Mi compañera de cuarto, que estudia psicología, me aplicó un test de personalidad y resulté neurótica. Desde entonces me he sentido muy molesta como para ir a clases."

"El año pasado, mientras realizaba una investigación usted aplicó a nuestros empleados un nuevo test de personalidad. Quisiéramos tener los resultados en sus expedientes."

Estos comentarios no son imaginarios; se basan en incidentes reales y cualquier psicólogo podría ampliar la lista. Ilustran abusos o malas interpretaciones de los instrumentos de medición psicológicos que podrían restarles todo valor o lastimar al individuo. Como cualquier instrumento científico o herramienta de precisión, las pruebas deben utilizarse correctamente para que sean eficaces. En manos de un usuario poco escrupuloso o bien intencionado pero ignorante pueden causar un grave daño. Hay dos razones principales para controlar su uso: (a) garantizar que sean aplicadas por un examinador calificado y que los resultados se empleen apropiadamente; y (b) impedir una familiaridad general con su contenido, ya que ello invalidaría el instrumento.

Examinador calificado. La necesidad de un examinador calificado se vuelve evidente en cada uno de los tres aspectos principales de la situación de prueba: la selección del test, su aplicación y calificación, y la interpretación de los resultados. Los tests no pueden elegirse como se escoge una podadora de un catálogo. No pueden evaluarse por el nombre, el autor u otras señas de identificación. Para estar seguros, no se requiere de entrenamiento psicológico al considerar factores como el costo, el volumen o la facilidad de transportación de los materiales de la prueba, el tiempo requerido para resolverla y la facilidad, así como la rapidez para calificarla. Por lo general, es posible obtener de un catálogo de tests la información sobre esos puntos prácticos que desde luego deben ser considerados al planear un programa de evaluación. Sin embargo, para que una prueba cumpla su función resulta imprescindible evaluar sus méritos técnicos en términos de sus características de validez, confiabilidad, grado de dificultad y normas. Sólo así es posible que los usuarios determinen qué tan adecuado es un instrumento para un propósito en particular o para las personas a las que planean aplicarlo.

Al hablar de la estandarización de las pruebas señalamos ya la importancia de contar con un examinador capacitado. Para que las puntuaciones que obtienen diferentes examinadores sean comparables o para evaluar los resultados de un individuo en tér-

minos de las normas publicadas es necesario percatarse de la importancia de seguir con precisión las instrucciones y de familiarizarse con ellas. También resulta fundamental el control cuidadoso de las condiciones de aplicación. De modo similar, la calificación incorrecta o inexacta puede inutilizar el resultado. Sin los procedimientos adecuados de supervisión, es mucho más probable que ocurran errores de calificación de lo que la gente cree.

La interpretación adecuada de los resultados requiere una comprensión cabal de la prueba, del sujeto que la presentó y de las condiciones en que fue aplicada. Sólo es posible determinar con objetividad lo que se mide si se hace referencia a los procedimientos que la validaron. También es pertinente contar con otra clase de información relativa a la confiabilidad, la naturaleza del grupo con el que se establecieron las normas, etc. Al interpretar los resultados es importante contar con algunos antecedentes del examinado. Distintas personas pueden obtener la misma puntuación por razones muy diferentes, por lo que las conclusiones a las que se llegue también deberían ser disímiles. Por último, también debe prestarse atención a factores especiales que pudieran haber afectado una calificación, como algunas condiciones inusuales de aplicación, el estado emocional o físico del examinado y su grado de experiencia con las pruebas.

La función del usuario. Durante los ochenta y los noventa el reconocimiento de la importante función del usuario constituyó un avance significativo en el campo de las pruebas psicológicas (Anastasi, 1990b). En este contexto, el usuario es cualquiera que utiliza los resultados de una prueba como fuente de información para tomar decisiones prácticas y puede ser, pero no necesariamente, el mismo que la aplica y la califica. Como ejemplos citemos a los maestros, consejeros, administradores de sistemas escolares o de personal en la industria o el gobierno. La mayor parte de las críticas no se dirige a los rasgos intrínsecos a las pruebas, sino al destino que usuarios mal calificados dan a los resultados. El deseo de encontrar atajos, respuestas rápidas y soluciones rutinarias simples para problemas complejos da lugar a algunos abusos. La presión temporal de una sobrecarga de trabajo puede fomentar tales recursos; sin embargo, es probable que la causa más frecuente sea el conocimiento insuficiente o inadecuado de las pruebas (Eyde, Moreland, Robertson, Primoff y Most, 1988; Moreland, Eyde, Robertson, Primoff y Most, 1995; Tyler y Miller, 1986).

En los Estados Unidos, comisiones especiales de organizaciones profesionales trabajan en conjunto con los editores de las pruebas para prevenir su mal uso. Un ejemplo notable es el proyecto del *Test User Qualifications Working Group* ("Grupo de Trabajo para la Certificación de los Usuarios de las Pruebas"), conocido por las siglas TUQWoG (Eyde *et al.*, 1988), cuya meta principal es el desarrollo de una base de datos empíricos de las condiciones esenciales que tienen que cumplir los usuarios de las pruebas y que los editores deben incluir en sus formas de certificación para permitir su adquisición. Luego de una investigación llevada a cabo durante cinco años en ese país, el proyecto TUQWoG formó una impresionante base de datos. Algunos editores ya han empezado a emplear los resultados en las formas de certificación del comprador. Más adelante se formó otro grupo con el propósito de utilizar la base de datos del TUQWoG para elaborar directrices y materiales de capacitación para los usuarios. El primer producto de este nuevo grupo, conocido como TUTWoG, Test User Training Working Group (Grupo de

Trabajo para la Capacitación de los Usuarios de las pruebas), es un libro que hace un recuento de los abusos más comunes con el propósito de prevenirlos (Eyde *et al.*, 1993). Los casos se basan en casos reales de abusos observados en diversos medios y que reveló una encuesta realizada para el proyecto. Moreland *et al.* (1995) presentan un resumen.

Seguridad del contenido de la prueba y comunicación de la información pertinente. Si una persona memorizara las respuestas correctas de una prueba de ceguera al color, la prueba quedaría totalmente invalidada dado que ya no podría ser una medida de su visión cromática. Es evidente que el contenido de las pruebas debe ser restringido para impedir los intentos por falsear los resultados; sin embargo, en otros casos el efecto de la familiaridad puede ser menos evidente o la prueba resultar invalidada de buena fe por personas mal informadas; por ejemplo, un maestro puede hacer que sus alumnos resuelvan problemas muy parecidos a los que presenta un test de inteligencia con el propósito de que "los niños estén preparados para presentarlo". Semejante actitud es simplemente un remanente del procedimiento que suele seguirse al preparar a los alumnos para un examen escolar; pero cuando se aplica a un test de inteligencia es probable que eleve las puntuaciones sin afectar de manera apreciable el área más amplia de conducta que la prueba pretende medir. En tales condiciones, se reduce la validez de la prueba como instrumento predictivo o de diagnóstico.

Garantizar la seguridad del contenido de una prueba no tiene por qué interferir con la comunicación de la información pertinente a las personas evaluadas, los profesionales interesados y el público en general, pues tal comunicación cumple varios objetivos. Primero, tiende a desvanecer cualquier velo de misterio asociado con el uso de las pruebas y en consecuencia ayuda a corregir los conceptos erróneos que prevalecen acerca de su propósito y el significado de sus resultados; para esto, algunos de los editores más importantes distribuyen folletos informativos. El segundo tiene que ver con los procedimientos técnicos seguidos al elaborar y evaluar los instrumentos; esta información ofrece datos importantes acerca de la confiabilidad, la validez y otras propiedades psicométricas del instrumento y, por lo general, se incluye en el manual técnico preparado para cada prueba y está disponible para cualquier persona interesada.

El tercer objetivo de la información consiste en familiarizar a las personas evaluadas con los procedimientos de la prueba, disminuir la ansiedad y lograr que cada una haga su mejor esfuerzo. Para estos fines se han preparado varios folletos explicativos, algunos de naturaleza general y otros para herramientas específicas como el Test de Evaluación Escolar de la Junta Universitaria (*College Board's Scholastic Assessment Test*), materiales que analizaremos en una sección posterior del capítulo. El cuarto objetivo, muy importante, es la retroalimentación que se brinda al examinado sobre su desempeño en la prueba. Los psicólogos han prestado una atención considerable a las formas más útiles y significativas de entregar esa información en diferentes contextos. En los capítulos 17 y 18 examinaremos los procedimientos apropiados.

La difusión de la información acerca de las pruebas es de gran importancia, y existen maneras útiles pero también dañinas de llevar a cabo esta tarea. Un ejemplo se encuentra en los precipitados intentos legislativos realizados en los Estados Unidos por introducir controles gubernamentales tanto a nivel estatal como federal (Bersoff, 1981, 1983; B. Lerner, 1980b). A finales de los setenta, fueron aprobadas leyes estatales que

regulan la divulgación de la información sobre las pruebas en California y en Nueva York. La de este último estado, que es la más extrema, requiere una divulgación estricta de las preguntas y respuestas de las pruebas aplicadas en todos los programas de evaluación a gran escala para la admisión a instituciones de educación superior.

Como semejante requisito supone la preparación de una nueva versión de cada prueba en cada ocasión que se aplica, puede tener varios efectos adversos como la disminución en las fechas de aplicación disponibles durante el año, el aumento en las cuotas que debe pagar el solicitante y la disminución en la calidad de los procedimientos de control al elaborar la prueba y al igualar las puntuaciones de las pruebas aplicadas en diferentes momentos. También es digno de mención que son muy pocas las personas que aprovechan la oportunidad que la legislación sobre divulgación les brinda, y que dicha divulgación no aumenta de manera significativa el desempeño en la segunda aplicación con otra forma de examen (Stricker, 1984). Las metas que impulsaron la promulgación de las leyes se alcanzan de mejor manera, y sin los nocivos efectos colaterales, si se fortalecen los procedimientos disponibles para comunicar la información de las pruebas.

APLICACIÓN DE LA PRUEBA

El fundamento de las pruebas es que pueden generalizar la muestra de conducta observada en la situación de prueba al comportamiento manifestado en otras situaciones. El resultado de una prueba debería ayudarnos a predecir cómo se sentirá y actuará el cliente fuera de la clínica, cuál será el desempeño académico del estudiante y cuál el desempeño laboral de un solicitante. Todas las influencias en la situación de prueba constituyen un error de varianza y reducen su validez. Por eso es tan importante identificar cualquiera que se relacione con la prueba y que pueda limitar o afectar la generalización de sus resultados.

Podríamos dedicar todo un volumen al análisis de los procedimientos deseables de aplicación de las pruebas, pero tal estudio escapa a los propósitos de este libro. Además, resulta más práctico adquirir dichas técnicas en medios específicos ya que, por lo general, no hay quien esté interesado en todas las formas de aplicación, del examen de infantes a las pruebas clínicas de pacientes psicóticos o a la aplicación de programas masivos de evaluación para personal militar. En consecuencia, nuestro análisis se orientará principalmente a los fundamentos de la aplicación de pruebas en lugar de abordar preguntas concretas sobre su puesta en práctica. Un excelente ejemplo de ello puede encontrarse en la concienzuda revisión de la evaluación individual infantil hecha por Sattler (1988, capítulo 5).

Preparativos previos a la aplicación. El requisito más importante de un buen procedimiento de aplicación son los preparativos. Durante la aplicación de la prueba no puede haber emergencias, por lo que tienen que hacerse esfuerzos especiales para anticiparlas e impedirlas. Sólo así se garantiza la uniformidad del procedimiento.

La preparación para la sesión de aplicación adopta muchas formas. En la mayor parte de las pruebas individuales resulta esencial memorizar las instrucciones verbales exactas. Incluso en una prueba de aplicación grupal en la que se leen las instrucciones

a los examinados, familiarizarse con el material previene los errores y dudas durante la lectura y permite un estilo más natural e informal durante la aplicación. Otro paso preliminar importante es la preparación de los materiales, que en las pruebas individuales y especialmente en las de desempeño consiste en disponer todo lo necesario para facilitar su uso con un mínimo de búsqueda o tropiezos. Es conveniente que los materiales se coloquen en un mueble cercano a la mesa en la que se realizará la aplicación de modo que estén al alcance del examinador pero que no distraigan al examinado. Cuando se emplean aparatos complejos puede ser necesario vigilarlos y calibrarlos de manera periódica. En las pruebas de aplicación grupal, todos los cuadernillos, las hojas de respuesta, los lápices especiales, o cualquier otro material, deben ser cuidadosamente contados, verificados y arreglados antes del día de la aplicación de la prueba.

Otro requisito importante, tanto en las pruebas de aplicación individual como en las de grupo, es la familiaridad absoluta con el procedimiento de aplicación. Para las pruebas individuales, suele ser esencial recibir una capacitación supervisada en la aplicación de la prueba. Dependiendo de la naturaleza del instrumento y de las personas examinadas; la capacitación puede requerir desde unas cuantas sesiones de demostración y práctica hasta más de un año de instrucción. Para las pruebas de grupo, en especial en los proyectos a gran escala, la preparación puede incluir instrucciones previas a los examinadores y ayudantes, de modo que todos estén completamente informados sobre las funciones que debe realizar. Por lo general, el examinador lee las instrucciones, se ocupa de llevar el tiempo y está a cargo del grupo en el salón. Los ayudantes entregan y recogen los materiales, se aseguran de que se sigan las instrucciones, responden las preguntas de los examinados dentro de los límites especificados en el manual e impiden las copias.

Condiciones de aplicación. El procedimiento estandarizado se aplica no sólo a las instrucciones verbales, el tiempo y otros aspectos de la prueba, sino también al ambiente. Es necesario elegir un salón adecuado para el examen, el cual debe estar libre de ruidos y distracciones y ofrecer a los examinados condiciones adecuadas de iluminación, ventilación, asientos y espacio de trabajo. También deben tomarse precauciones para prevenir interrupciones durante la aplicación, por lo que es conveniente colocar en la puerta un cartel que indique que se está aplicando la prueba y asegurarse de que todo el personal se haya enterado de que la señal significa que nadie puede entrar bajo ninguna circunstancia. En las pruebas colectivas, puede ser necesario cerrar la puerta o poner a un ayudante afuera para impedir la entrada de los rezagados.

Es importante darse cuenta de las condiciones en que se realiza la prueba ya que éstas pueden influir en sus resultados. Incluso aspectos que parecen menores pueden alterar de manera apreciable el desempeño; por ejemplo, el uso de pupitres fijos o bien de sillas con paleta para el brazo demostró ser significativo en un proyecto de aplicación grupal con estudiantes de secundaria, pues el grupo que utilizó pupitres tendía a obtener mayores calificaciones (T. L: Kelley, 1943; Traxler y Hilkert, 1942). También hay evidencias que demuestran que la hoja de respuesta que se emplee puede influir en los resultados (F. O. Bell, Hoff y Hoyt, 1964). A veces, los examinadores utilizan en sus pruebas colectivas hojas de respuesta diferentes de las que se emplearon en la muestra de estandarización, lo que se debe al establecimiento de oficinas independientes de

calificación de pruebas y procesamiento de datos que entregan sus propias hojas de respuesta, las que pueden ser calificadas por máquinas. Dado que se carece de una verificación empírica, no es posible asumir que las hojas son equivalentes. Al examinar a niños de grados inferiores al quinto, el uso de *cualquier* hoja diferente puede disminuir de forma notable sus calificaciones (Cashen y Ramseyer, 1969; Ramseyer y Cashen, 1971), así que en esos niveles suele ser preferible hacer que marquen las respuestas en el propio cuadernillo de la prueba.

Todavía más significativas a cualquier edad son las diferencias entre la aplicación computarizada o de lápiz y papel de la misma prueba, por ello se ha dedicado considerable atención al efecto que tiene sobre las normas, la confiabilidad y la validez en relación con la naturaleza de la prueba y la población examinada. Lo anterior ha conducido a la formulación de lineamientos profesionales para que los usuarios decidan qué tan comparables son los resultados de las dos aplicaciones (Butcher, 1987; Hofer y Green, 1985).

Se ha demostrado que muchas otras condiciones sutiles afectan el desempeño tanto en los tests de habilidad como en los de personalidad. Que el examinador sea extraño o conocido para el examinado puede suponer una diferencia significativa en los resultados (Sacks, 1952; Tsudzuki, Hata y Kuze, 1957). En otro estudio se comprobó que los modales y la conducta del examinador (sonreír, asentir con la cabeza y hacer comentarios como "bien" o "perfecto"), tienen un efecto decisivo en los resultados (Wickes, 1956). Al aplicar una técnica proyectiva que requiere que el examinado escriba historias que se ajusten a ciertas imágenes, se descubrió que la presencia del examinador en la habitación tendía a inhibir la inclusión de contenido emocional en las historias (Bernstein, 1956). En la aplicación de una prueba de mecanografía, los solicitantes escribían a una tasa significativamente más alta cuando se les probaba solos que cuando el examen se hacía en grupos de dos o más personas (Kirchner, 1966).

Podríamos multiplicar con facilidad los ejemplos, hecho que tiene tres implicaciones. Primero, siga con minucioso detalle los procedimientos estandarizados. Es responsabilidad del autor de la prueba y del editor explicar los procedimientos de manera clara y completa en el manual de la prueba. Segundo, lleve registro de cualquier condición inusual que tenga lugar durante la aplicación, aunque sea menor. Tercero, al interpretar los resultados, tome en consideración las condiciones en las que se llevó a cabo. En la evaluación a fondo de un sujeto, el examinador experimentado ocasionalmente puede apartarse del procedimiento estandarizado para obtener información adicional por razones especiales. En esos casos, ya no se puede interpretar los resultados en términos de las normas de la prueba. En tales circunstancias, los estímulos de la prueba se utilizan únicamente para una exploración cualitativa, y las respuestas tendrían que tratarse de la misma manera que cualquier otra observación informal de la conducta o que los datos obtenidos en una entrevista.

Presentación de la prueba: rapport y orientación del examinado. Al aplicar una prueba, se entiende por *rapport* los esfuerzos del examinador por despertar el interés del examinado, lograr su cooperación y animarlo a responder de manera apropiada a los objetivos del instrumento. En los tests de habilidad, el objetivo requiere una concentración cuidadosa en las tareas presentadas y hacer el mejor esfuerzo por un buen desempe-

ño. En los inventarios autodescriptivos de personalidad, el objetivo es obtener respuestas francas y honestas a las preguntas sobre las conductas cotidianas; en ciertas técnicas proyectivas, se requiere de un informe completo de las asociaciones evocadas por los estímulos sin censura ni manipulación del contenido. Otras pruebas exigen otras aproximaciones, pero en todos los casos, el examinador se esfuerza por motivar al examinado a seguir las instrucciones de manera tan completa y concienzuda como le sea posible.

La capacitación de los examinadores incluye la adquisición de técnicas para el establecimiento de *rapport*, así como de otras que están relacionadas más directamente con la aplicación de la prueba. En el establecimiento del *rapport*, como en otros procedimientos de la prueba, resulta esencial la uniformidad de las condiciones para que los resultados sean comparables. Si una niña recibe un premio cada vez que resuelve un problema de la prueba, su desempeño no puede ser comparado directamente con las normas o con el desempeño de otros niños que sólo son motivados con incentivos o alabanzas verbales. Cualquier desviación de las condiciones motivacionales normales de una prueba tiene que anotarse y considerarse al interpretar la ejecución.

Aunque el *rapport* se establece más plenamente en las pruebas individuales, en las de grupo también es posible emprender acciones para motivar a los examinados y disminuir su ansiedad. Las técnicas varían con la naturaleza de la prueba, la edad y otras características del examinado. Cuando se trabaja con preescolares, deben considerarse factores especiales como la timidez ante los desconocidos, la disposición a distraerse y el negativismo. Un estilo amistoso, cariñoso y relajado de parte del examinador ayuda a darles confianza. El niño tímido y retraído necesita más tiempo para familiarizarse con los alrededores, por lo que es mejor que el examinador no se extienda demasiado al principio, sino que espere a que el niño esté listo para hacer el primer contacto. Los periodos de prueba deben ser breves y las tareas variadas e interesantes para el niño; tienen que presentarse como un juego, y antes de introducir una tarea nueva hay que despertar su curiosidad. A esta edad se requiere cierta flexibilidad de procedimiento por la posibilidad de negativas, pérdida de interés y otras manifestaciones de negativismo.

Los niños en los dos o tres primeros grados de la escuela elemental presentan muchos de los problemas observados en los preescolares, por lo que el método del juego sigue siendo la forma más eficaz de despertar su interés en la prueba. Los mayores pueden ser motivados si se apela al espíritu de competencia y al deseo de hacer un buen trabajo; sin embargo, al evaluar a niños cuyos antecedentes educativos los ponen en desventaja o que provienen de culturas diferentes, el examinador no puede suponer que estarán motivados para destacar en las tareas académicas en el mismo grado que los niños de la muestra de estandarización. En los capítulos 9, 12 y 18 veremos éste y otros problemas relacionados con la evaluación de sujetos con experiencias disímiles.

Es probable encontrar problemas motivacionales especiales al evaluar a individuos con perturbaciones emocionales, prisioneros y delincuentes juveniles, que posiblemente manifiesten actitudes desfavorables como suspicacia, inseguridad, temor o indiferencia cínica, en especial cuando son evaluados en un centro de reclusión. También es probable que ciertas peculiaridades de sus experiencias influyan en su desempeño de manera nociva; por ejemplo, como resultado de los fracasos y las frustraciones sufridos en la escuela pueden haber desarrollado sentimientos de hostilidad e inferioridad hacia las tareas académicas, que se parecen a las pruebas. El examinador

experimentado debe hacer esfuerzos especiales por establecer el *rapport* en tales condiciones. En cualquier caso, debe ser sensible a esas dificultades especiales y tomarlas en consideración al interpretar y explicar el desempeño en la prueba.

Al evaluar a niños escolares o a adultos debe recordarse que cada prueba representa una amenaza implícita para el prestigio del individuo, por lo que es necesario tranquilizarlo desde el inicio; por ejemplo, es útil explicarle que no se espera que nadie termine o responda correctamente todos los reactivos, pues, de otra manera, el examinado experimentaría una sensación de fracaso al avanzar en los reactivos más difíciles o al ver que no es capaz de terminar algún segmento en el tiempo permitido.

Dado que lo inesperado y lo desconocido suelen producir ansiedad, también es deseable eliminar, hasta donde resulte posible, las sorpresas en la situación de prueba. Aunque muchas pruebas colectivas incluyen una explicación preliminar que el examinador lee ante el grupo, un procedimiento aún mejor consiste en proporcionar con antelación a cada examinado materiales que expliquen el propósito y la naturaleza del instrumento, le ofrezcan sugerencias generales sobre cómo presentarla y que contengan algunos reactivos de muestra. Por lo general, quienes participan en programas de evaluación a gran escala tienen acceso a los manuales explicativos, como los llevados a cabo por la Junta Universitaria (*College Board*).

La prueba para adultos implica otros problemas, pues a diferencia de los escolares es poco probable que se esfuercen en una tarea simplemente porque les es asignada. Por ello se hace más importante "venderles" el propósito de la prueba, aunque los estudiantes de educación media y superior también responden a ese llamado. Habitualmente, es posible lograr la cooperación de los examinados al convencerlos de que les conviene obtener una puntuación válida, es decir, un resultado que indique correctamente lo que pueden hacer en lugar de sobrestimar o subestimar sus habilidades. La mayoría de la gente puede entender que una decisión incorrecta, tomada a partir del resultado no válido de la prueba, puede significarle fracasos, pérdida de tiempo y frustraciones. Como los sujetos se dan cuenta de que son ellos los que más tienen que perder, este sistema puede servir no sólo para motivarlos a hacer su mejor esfuerzo en los tests de habilidad, sino también para reducir los engaños y fomentar el reporte honesto en los inventarios de personalidad. Por supuesto, a nadie le conviene ser admitido en un curso para el que no tiene las habilidades o el conocimiento requeridos o ser asignado a un puesto que no puede desempeñar o que encuentra desagradable.

EXAMINADOR Y VARIABLES SITUACIONALES

Periódicamente se publican estudios de gran alcance sobre los efectos que el examinador y las variables situacionales tienen en los resultados de las pruebas (Lutey y Copeland, 1982; Masling, 1960; S. B. Sarason, 1954; Sattler, 1970, 1988; Sattler y Theye, 1967). Aunque se han descubierto algunos efectos en las pruebas objetivas de grupo, la mayor parte de los datos proviene de técnicas proyectivas o de tests de inteligencia. Es más probable que esas variables extrañas operen con estímulos ambiguos y no estructurados, así como con tareas difíciles y novedosas, que con funciones definidas con claridad y bien aprendidas. En general, los niños son más susceptibles que los adultos a los efectos de las variables situacionales y del aplicador, por lo que la función

de éste resulta especialmente importante al evaluar a preescolares. La probabilidad de que estas condiciones tengan algún influjo también es mayor en las personas inseguras o con trastornos emocionales de cualquier edad.

Se ha investigado la relación entre el desempeño en tests de inteligencia y técnicas proyectivas aplicadas individualmente con muchas variables del examinador, entre las que se incluyen edad, sexo, raza, posición profesional o socioeconómica, capacitación y experiencia, características de personalidad y apariencia. Aunque se han encontrado varias relaciones significativas, los resultados suelen ser erróneos o no concluyentes porque el diseño experimental no logró controlar o aislar la influencia de diferentes características del aplicador o del examinado, lo que supone la posible confusión de los efectos de dos o más variables.

Lo que se ha demostrado con mayor claridad es que la conducta del aplicador antes y durante la prueba puede alterar los resultados; por ejemplo, en investigaciones controladas se ha descubierto diferencias notables en el desempeño en un test de inteligencia como resultado de una relación interpersonal "cálida" frente a una "fría" entre aplicador y examinado, o un estilo del aplicador rígido y reservado frente a uno más natural (Exner, 1966; Masling, 1959). Más aún, puede haber interacciones significativas entre las singularidades del aplicador y examinado en el sentido de que las mismas características del aplicador, o su estilo de aplicar la prueba, pueden tener un efecto muy diferente en distintos examinados en función de la propia personalidad del examinado. Interacciones similares pueden ocurrir con las variables de la tarea, como la naturaleza de la prueba, el propósito de su aplicación y las instrucciones dadas a los evaluados. Dyer (1973) ha agregado otras variables a esta lista, y llama la atención sobre la posible influencia de las diversas percepciones que aplicadores y examinados tienen de las funciones y objetivos de la prueba.

Las expectativas del aplicador constituyen otra forma en que éste puede afectar sin quererlo las respuestas del examinado. Se trata simplemente de un caso especial de la profecía autorrealizada (Harris y Rosenthal, 1985; R. Rosenthal, 1966; R Rosenthal y Rosnow, 1969). Un ejemplo se encuentra en un experimento realizado con el test de Rorschach (Masling, 1965). Los aplicadores fueron 14 estudiantes graduados que se ofrecieron como voluntarios; a siete de ellos se les dijo, entre otras cosas, que los aplicadores experimentados provocaban más respuestas humanas que animales, mientras que a los otros siete se les informó lo contrario. En tales condiciones, los dos grupos obtuvieron de sus examinados proporciones significativamente diferentes de respuestas animales o humanas, y esas diferencias ocurrieron a pesar de que ni los aplicadores ni los examinados dijeron estar conscientes de ninguna influencia. Más aún, las grabaciones de las sesiones no revelaron evidencias de influencias verbales por parte de ningún aplicador. Según parece las expectativas de los aplicadores operaron con sutiles claves posturales y faciales a las que respondieron los examinados.

Además del aplicador, otros aspectos de la situación de prueba pueden influir significativamente en el desempeño; por ejemplo, a menudo se evalúa a los reclutas al poco tiempo de su inducción, durante un periodo de intenso reajuste a una situación desconocida y estresante. En una investigación destinada a probar el efecto de aclimatarse a dicha situación sobre el desempeño en la prueba, se aplicó la Batería de Clasificación de la Marina (*Navy Classification Battery*) a 2 724 reclutas en su noveno día

en el Centro de Entrenamiento Naval (L. V. Gordon y Alf, 1960). Al comparar sus resultados con los de 2 180 reclutas probados en el momento habitual, al tercer día en el centro, el primer grupo obtuvo calificaciones superiores en todos los subtests de la batería.

Las actividades que realizan los sujetos justo antes de la prueba también tienen un efecto en su desempeño, en particular si producen perturbación emocional, fatiga u otras condiciones desventajosas. En una investigación realizada con niños de tercer y cuarto grado se hallaron evidencias de que la actividad que había tenido lugar antes en el aula influía en el CI que obtenían en la Prueba de Dibujo de un Hombre (McCarthy, 1944). En una ocasión, los estudiantes habían estado ocupados en la redacción de una composición sobre "Lo mejor que me ha sucedido", mientras que en la segunda escribían sobre "Lo peor que me ha pasado". El CI de la segunda prueba, después de lo que puede haber sido una experiencia depresiva, era en promedio cuatro o cinco puntos inferior al de la primera prueba. Estos descubrimientos fueron corroborados en otra investigación diseñada en concreto para determinar el efecto de la experiencia inmediatamente precedente sobre la Prueba de Dibujo de un Hombre (*Draw-a-Man Test*) (Reichenberg-Hackett, 1953). En este estudio, los niños que habían pasado por la experiencia gratificante de resolver un acertijo interesante y recibir por recompensa dulces y juguetes, mostraron mejores puntuaciones que quienes habían tenido una experiencia neutral o menos agradable. W. E. Davis (1969a, 1969b) obtuvo resultados similares con estudiantes universitarios. El desempeño en una prueba de razonamiento aritmético fue significativamente inferior cuando la precedía una experiencia fallida en una prueba de comprensión verbal que en un grupo de control al que no se aplicó dicha prueba o en otro que había presentado una prueba estándar de comprensión verbal en condiciones ordinarias.

Diversos estudios se han interesado por los efectos de la retroalimentación en los resultados de la prueba en la ejecución subsecuente del individuo. En una investigación muy bien diseñada con estudiantes de séptimo grado, Bridgeman (1974) encontró que el desempeño en una prueba subsecuente era bastante mejor cuando se recibía retroalimentación de "éxito" que cuando era de "fracaso" en una prueba inicial en la que la ejecución de ambos grupos había sido igualmente buena. La retroalimentación motivacional funciona sobre todo mediante las metas que los participantes se imponen para su desempeño posterior, por lo que es otro ejemplo de la profecía que se autorrealiza; sin embargo, no debe confundirse esta retroalimentación motivacional con la correctiva, con la que se informa al individuo de los reactivos específicos que contestó mal y recibe educación de regularización. En estas condiciones, es mucho más probable que la retroalimentación mejore la ejecución de quienes habían obtenido un mal resultado.

Los ejemplos citados ilustran la gran diversidad de variables relacionadas con las pruebas que pueden alterar los resultados. En la mayor parte de los programas de evaluación bien aplicados, la influencia de esas variables es insignificante para todo propósito práctico; no obstante, el examinador experimentado siempre está en guardia para detectar la operación de dichas variables y disminuir su influjo. Cuando las circunstancias no permiten el control de estas condiciones, resulta necesario restringir las conclusiones derivadas del desempeño en la prueba.

EL PUNTO DE VISTA DEL EXAMINADO

Ansiedad ante la prueba. Entre las primeras investigaciones sobre las reacciones de los evaluados ante la situación de prueba se encuentran las que estudian la ansiedad que produce la prueba. Es indudable que su notoriedad y sus efectos nocivos sobre el desempeño estimularon el interés por esta información. En la aplicación de las pruebas, muchas de las prácticas destinadas a lograr el *rapport* también reducen la ansiedad, lo mismo que los procedimientos que tienden a disminuir las sorpresas y la extrañeza de la situación de prueba y a tranquilizar y animar al examinado. El estilo del aplicador y una sesión bien organizada contribuyen al mismo fin.

Las diferencias individuales en cuanto a la ansiedad que causan las pruebas han sido estudiadas lo mismo en escolares que en universitarios (Gaudry y Spielberger, 1974; Hagtvet y Johnsen, 1992; I. G. Sarason, 1980; Spielberger, 1972). Buena parte de esta investigación fue iniciada por S. B. Sarason y sus colaboradores en Yale (Sarason, Davidson, Lighthall, Waite y Ruebush, 1960). El primer paso consistió en formular un cuestionario para evaluar las actitudes del individuo hacia la prueba; por ejemplo, la forma para los niños contenía reactivos como los siguientes:

¿Te preocupas mucho antes de presentar una prueba?

¿Sientes que tu corazón empieza a latir más aprisa cuando la maestra dice que va a averiguar qué tanto has aprendido?

Cuando estás presentando una prueba ¿piensas que no lo estás haciendo bien?

El principal interés es el descubrimiento de que tanto los resultados de las pruebas de rendimiento académico como los de los tests de inteligencia arrojaron correlaciones negativas con la ansiedad producida por las pruebas. En los estudiantes universitarios se han encontrado correlaciones similares (I. G. Sarason, 1961). Del mismo modo, estudios longitudinales revelaron una relación inversa entre los cambios en el grado de ansiedad y los cambios en el desempeño en tests de inteligencia o de aprovechamiento (K. T. Hill y S. B. Sarason, 1966; S. B. Sarason, K. T. Hill y Zimbardo, 1964).

Por supuesto, estos descubrimientos no indican la presencia de relaciones causales. Es posible que los estudiantes se sientan ansiosos con las pruebas porque suelen tener un mal desempeño al realizarlas y por ende han pasado por frustraciones y fracasos. En apoyo a esta interpretación está el hecho de que en los subgrupos de mayores puntuaciones en los tests de inteligencia desaparece la correlación negativa entre el grado de ansiedad y el desempeño (Denny, 1966; Feldhusen y Klausmeier, 1962). Por otro lado, se cuenta con evidencias de que al menos parte de esta relación proviene de los efectos nocivos de la ansiedad sobre el desempeño en las pruebas. En un estudio (Waite, Sarason, Lighthall y Davidson, 1958), niños con diferentes grados de ansiedad (elevada y baja), pero con resultados similares en los tests de inteligencia, hicieron varios ensayos en una tarea de aprendizaje. Aunque al principio su ejecución en la prueba de aprendizaje era igual, el grupo con menor ansiedad mostró una mejora significativamente mayor que el grupo ansioso.

Varios investigadores han comparado la ejecución en las pruebas en condiciones diseñadas para producir un estado "ansioso" o uno "relajado"; por ejemplo, Mandler y

Sarason (1952) descubrieron que instrucciones que se referían al ego (como decir a los examinados que se esperaba que todos terminaran en el tiempo permitido) tenían un efecto benéfico sobre la ejecución de los más tranquilos, pero uno nocivo en los ansiosos. Otros estudios también han encontrado una relación entre las condiciones de la prueba y las características individuales como el nivel de ansiedad y la motivación para el logro (Lawrence, 1962, Paul y Eriksen, 1964). Parece probable que la relación entre ansiedad y ejecución no sea lineal, es decir, que una poca de ansiedad sea benéfica y que mucha sea nociva. Los individuos que habitualmente son poco ansiosos se benefician de las condiciones de la prueba que generan cierta ansiedad, mientras que los que suelen ser presa del nerviosismo trabajan mejor en condiciones más relajadas.

No cabe duda de que una ansiedad elevada crónica ejerce un efecto nocivo sobre el aprendizaje académico y el desarrollo intelectual. La ansiedad interfiere con la adquisición y la recuperación de la información (Hagtvet y Johnsen, 1992). Sin embargo, es necesario distinguir este efecto del impacto que la ansiedad tiene sobre la prueba (es decir, la medida en que la ansiedad que produce hace que el desempeño del individuo sea poco representativo de su desempeño en otras situaciones), que es lo que constituye el objeto de nuestro estudio. Debido a la presión competitiva que experimentan los estudiantes cuyo ingreso a la universidad está próximo, se ha argumentado que el desempeño en las pruebas de admisión puede resultar muy afectado por la ansiedad que causan. En una investigación cuidadosa y bien diseñada, French (1962) comparó el desempeño de estudiantes de bachillerato en un examen aplicado como parte de la aplicación regular del Test de Aptitud Académica (*Scholastic Aptitude Test*) con su ejecución en una forma paralela de esa prueba aplicada en otro momento en condiciones "relajadas". Las instrucciones en este último caso especificaban que la prueba se aplicaba con propósitos de investigación y que los resultados no se enviarían a ninguna universidad. Ahora bien, éstos mostraron que la ejecución no fue peor durante la aplicación estándar que durante la relajada. Más aún, la validación de los puntajes de la prueba con las notas escolares no difería en las condiciones. Varias investigaciones recientes también han cuestionado el estereotipo común del estudiante ansioso por la prueba que sabe la materia pero que se "paraliza" al presentar el examen (véase Culler y Holahan, 1980). En esta investigación se descubrió que los estudiantes que calificaban más alto en una escala de ansiedad obtenían en promedio menores calificaciones y solían tener peores hábitos de estudio que los que habían calificado bajo.

La investigación sobre la naturaleza, la medición y el tratamiento de la ansiedad producida por los exámenes ha continuado a paso creciente (I. G. Sarason, 1980; Spielberger, Anton y Bedell, 1976; Spielberger, González y Fletcher, 1979; Spielberger, González, Taylor, Algaze y Anton, 1978; G. S. Tryon, 1980) y ha permitido identificar dos componentes importantes relacionados con la naturaleza de la ansiedad, la emocionalidad y la preocupación. El componente de emocionalidad comprende sentimientos y reacciones fisiológicas, como tensión y aumento del ritmo cardiaco. La preocupación, o componente cognoscitivo, incluye pensamientos negativos sobre sí mismo, como la expectativa de fracaso y el temor por sus consecuencias. Esos pensamientos desvían la atención de la conducta orientada a la tarea que la prueba demanda y, por consecuencia, trastornan el desempeño. Diversos inventarios de ansiedad miden ambos componentes, y aunque son de uso común en la investigación, hasta hace muy poco

sólo estaban disponibles en los informes de la bibliografía empírica. El Inventario de Ansiedad (*Test Anxiety Inventory*) elaborado por Spielberger y sus colaboradores es un ejemplo que explicamos en el capítulo 13 y que aparece en el apéndice A.

Se ha dedicado un considerable esfuerzo al desarrollo y la evaluación de métodos para el tratamiento de la ansiedad producida por las pruebas que incluyen varios procedimientos de terapia conductual (capítulo 17) para reducir el componente emocional. Los resultados han sido en general buenos, pero debido a las fallas metodológicas de los estudios de evaluación resulta difícil atribuir la mejoría a alguna técnica (G. S. Tryon, 1980). De hecho, este componente en el test de ansiedad tiende a disminuir del test al retest incluso en los grupos de control sin intervención terapéutica, así como en grupos de control especiales que recibieron una pseudoterapia creíble. Más aún, su reducción tuvo poco o ningún efecto sobre el nivel de desempeño.

Cuando el tratamiento se dirige a las reacciones cognoscitivas personales es más probable que mejore el desempeño tanto en las pruebas como en el trabajo escolar. La investigación disponible hasta ahora indica que los mejores resultados se obtienen al combinar programas de tratamiento para eliminar la emocionalidad y la preocupación así como la mejora de los hábitos de estudio. La ansiedad producida por los exámenes es un fenómeno complejo con causas múltiples, y la contribución relativa de cada causa varía con el individuo. Para que funcionen, los programas de tratamiento tendrían que adaptarse a las necesidades individuales. También debe reconocerse que esta ansiedad es sólo una manifestación de un conjunto más general de condiciones que reducen la eficacia del individuo para aprender.

Investigación amplia de las opiniones de los examinados. Aunque la ansiedad producida por los exámenes es una parte evidente e importante de la conducta de quien presenta una prueba, hay muchos otros elementos que pueden ser estudiados con provecho. Un libro editado en 1993 por Baruch Nevo y R. S. Jäger representa un esfuerzo notable por reunir la información disponible sobre las reacciones de los examinados a las pruebas en medios educativos, industriales, clínicos y de orientación. En los 15 capítulos redactados por investigadores reconocidos sobre diversos aspectos y aplicaciones de las pruebas se encuentran lo mismo informes de publicaciones internacionales sobre cada tema como los descubrimientos de los propios autores. El resultado es un intento serio y fundado por resolver cuestiones que hasta ahora han sido tratadas sobre todo en entornos periodísticos, políticos o legales. El libro funge así como correctivo para las opiniones posiblemente sesgadas y conflictivas sobre las pruebas que abundan en la actualidad; por ejemplo, el primer capítulo incluye 10 excelentes encuestas de opinión sobre las actitudes hacia las pruebas de una amplia gama de poblaciones. Los resultados revelan discrepancias entre las opiniones del público y las afirmaciones de algunos voceros muy publicitados pero poco representativos.

Los capítulos cubren numerosos temas. Algunos explican el desarrollo y el uso de cuestionarios de retroalimentación y las entrevistas de grupo para evaluar las actitudes de diferentes grupos de examinados hacia determinada prueba y sus percepciones sobre lo que ésta mide. En un capítulo comparó las opiniones de estudiantes sobre los exámenes escolares de ensayo y de opción múltiple, que fueron los favorecidos por los resultados. Algunos autores estudian las reacciones de los solicitantes de empleo hacia la

equidad de las pruebas y su relación con el trabajo. Como resultado de sus descubrimientos, varios capítulos sugieren formas de mejorar la aplicación y el ambiente de las pruebas. En conjunto, los capítulos abren a la exploración un área prometedora para buscar soluciones a algunos de los problemas sociales y prácticos de las pruebas, además de que brindan los medios para aumentar la comprensión recíproca de los usuarios de las pruebas y los examinados.

EFECTOS DEL ENTRENAMIENTO SOBRE EL DESEMPEÑO EN LA PRUEBA

Al evaluar los efectos del entrenamiento o la práctica sobre los resultados de las pruebas, una pregunta fundamental es si la mejora se limita a los reactivos que incluye o si se extienden al área más amplia de conducta que la prueba pretende evaluar (Anastasi, 1981a, 1981b). La respuesta a esta pregunta muestra la diferencia entre preparación y educación. Obviamente, cualquier experiencia educativa, formal o informal, dentro o fuera de la escuela, debería reflejarse en el desempeño en las pruebas que estudian los aspectos pertinentes del comportamiento. Dichas influencias no invalidan la prueba en la medida en que su resultado presenta una imagen exacta de la posición del individuo en las habilidades consideradas. Por supuesto, la diferencia es de grado. Las influencias no pueden clasificarse como restringidas o amplias sino que varían en su alcance, de las que afectan una única aplicación de una sola prueba, a las que alteran el desempeño en todos los reactivos de cierta clase y las que influyen en el desempeño del individuo en casi todas las actividades; sin embargo, desde el punto de vista de un examen eficaz, es posible hacer una distinción útil. Así, puede afirmarse que el resultado de la prueba sólo es inválido cuando una experiencia particular eleva la puntuación sin modificar en forma apreciable el área de conducta que pretende medir.

Preparación. Los efectos de la preparación sobre los resultados de las pruebas han sido objeto de una amplia investigación. Los psicólogos británicos realizaron varios estudios que se refieren en especial a los efectos de la práctica y la preparación sobre las pruebas que solían utilizarse para asignar a los niños de 11 años a diferentes escuelas secundarias (Yates *et al.*, 1953-1954). Como era de esperarse, el grado de mejoramiento dependía de la habilidad del examinado, de sus experiencias educativas, de la naturaleza de las pruebas y de la cantidad y clase de preparación recibida. Los individuos con antecedentes educativos deficientes tenían mayores probabilidades de beneficiarse de la preparación especial que quienes habían tenido mejor educación y ya estaban preparados para desempeñar un buen papel en las pruebas. También es evidente que entre más estrecho fuera el parecido entre el contenido de la prueba y el material empleado en la preparación, mayor sería la mejora en los resultados. Por otro lado, entre más se restrinja la instrucción al contenido de la prueba, menos probable es que la mejora se extienda al desempeño de criterio. Más aún, muchos estudios sobre la preparación han arrojado resultados ambiguos y difíciles de interpretar debido a serias deficiencias metodológicas (Anastasi, 1981a; Bond, 1989; Messick, 1980a), entre las cuales sobresale la incapacidad para emplear un grupo de control sin preparación que sea verdaderamente equiparable al grupo preparado; por ejemplo, los estudiantes que se inscriben en los programas comer-

ciales de preparación son autoseleccionados y su habilidad inicial, motivación y otras características personales que influyen en el desempeño de la prueba tienden a diferir de los del grupo de control. Además, en los diseños experimentales que emplean pretest y postest es difícil asegurar que los examinados estén igualmente motivados para hacer un buen trabajo en ambas ocasiones, lo que es cierto sobre todo si una prueba tiene una aplicación regular y la otra una especial con propósitos de práctica o de investigación.

La Junta Universitaria de Exámenes de Admisión a la Universidad (*College Entrance Examination Board*) se encuentra preocupada por la proliferación de cursos comerciales que ofrecen preparar a los estudiantes que aspiran a ingresar en las universidades. Para aclarar el tema, ha realizado diversos experimentos bien controlados para determinar el efecto de la preparación sobre el Test de Aptitud Académica (*Scholastic Aptitude Test*, SAT) y ha revisado los resultados de estudios similares realizados por investigadores independientes (Donlon, 1984; Messick, 1980a, 1981; Messick y Jungeblut, 1981). Los estudios cubren numerosos métodos de preparación e incluyen a estudiantes de escuelas públicas y privadas, así como de grupos minoritarios de áreas urbanas y rurales. La conclusión general es que es poco probable que el ejercicio intenso en reactivos similares a los del SAT produzca ganancias mayores que las que se observan cuando éste se aplica nuevamente luego de un año de instrucción regular.

También debemos decir que en sus procedimientos de elaboración de instrumentos, organizaciones como la Junta Universitaria (*College Board*) y el Consejo de Exámenes de Registro de Graduados (*Graduate Record Examination Board*) investigan la susceptibilidad a la preparación de nuevos reactivos (Evans y Pike, 1973; Powers, 1983; Powers y Swinton, 1984; Swinton y Powers, 1985). En las formas operacionales de la prueba no se conservan los reactivos cuya ejecución puede mejorar mucho con el ejercicio o instrucción a corto plazo y que tienen una naturaleza sumamente restringida. Un ejemplo evidente es el problema que requiere una solución sencilla y perspicaz que, una vez alcanzada, puede aplicarse tal cual a la solución de problemas similares. Cuando vuelven a aparecer, los problemas recuerdan la prueba más que las habilidades de solución de problemas. Otro ejemplo se encuentra en los reactivos complejos que utilizan material novedoso o desconocido y requieren instrucciones largas y complicadas (Powers, 1986).

En el sentido tradicional, la preparación pretende desarrollar habilidades muy restringidas que pueden ser de poca utilidad en las actividades cotidianas. De modo similar, la práctica de "enseñar la prueba" tiende a concentrarse en la muestra particular de habilidades y conocimientos que cubre más que en el área general de conocimientos que la prueba pretende evaluar. Las llamadas leyes de divulgación u "honestidad de las pruebas" que requieren la publicación general de las formas utilizadas luego de una sola aplicación también favorecen la concentración en las habilidades específicas de la prueba, cuya aplicabilidad es limitada. Por último, en la medida en que la preparación sólo está al alcance de algunos, tiende a introducir diferencias individuales en las habilidades específicas de presentación de exámenes, lo que reduce la capacidad de diagnóstico del instrumento.

Perfeccionamiento en las pruebas. A este respecto, también son notables los efectos del perfeccionamiento en las pruebas, o la mera práctica de presentarlas. En estudios con versiones de la misma prueba se observa una tendencia a que la segunda calificación sea más alta. Se ha informado de ganancias significativas en promedio al administrar formas alternas en sucesión inmediata o después de lapsos que van de un

día a tres años (Donlon, 1984; Droege, 1966; Peel, 1951, 1952). Se han obtenido resultados similares con niños normales y sobredotados, estudiantes de educación media y superior y con muestras de empleados. El manual de la prueba debe ofrecer datos sobre la distribución de ganancias esperadas en el retest con una forma paralela, y hay que descontarlas al interpretar los resultados.

Las ganancias no se limitan a las formas alternas, los que tienen mucha experiencia en la presentación de pruebas estandarizadas disfrutan de cierta ventaja sobre quienes presentan la prueba por primera vez (Millman, Bishop y Ebel, 1965; Rodger, 1936). Parte de esta ventaja se debe a que han superado la sensación inicial de extrañeza y a que han adquirido más confianza y mejores actitudes hacia las pruebas, pero también es resultado de cierto traslape en los contenidos y las funciones de muchas pruebas. La familiaridad con algunos reactivos comunes y la práctica en el uso de las hojas de respuesta objetiva también pueden mejorar ligeramente el desempeño. Al comparar los resultados de sujetos con distintos grados de experiencia en las pruebas es importante tomar en cuenta este perfeccionamiento. Para las pruebas aplicadas por computadora debe prestarse atención a la familiaridad del examinado con esta forma de administración (Hofer y Green, 1985).

Las condiciones de perfeccionamiento pueden ser equiparadas de manera eficaz por medio de una breve orientación y sesiones de práctica (Anastasi, 1981a; Wahlstrom y Boersman, 1968). Esta familiarización reduce los efectos de las diferencias previas en la experiencia con las pruebas. La disminución de las diferencias, que son particulares de la situación de prueba, debería permitir una evaluación más válida del área general de conducta que la prueba pretende medir. Tal planteamiento lo ilustra la publicación de la Junta Universitaria titulada *Taking the SAT I: Reasoning Test* ("La presentación del SAT I: Prueba de razonamiento"), un cuadernillo distribuido a todos los aspirantes a la educación universitaria que se registran para presentar esta prueba y que ofrece consejos para prepararla en forma eficaz, ilustra y explica los diferentes reactivos que incluye y reproduce una forma completa de la misma, con la sugerencia a los estudiantes de resolverla en las condiciones normales de tiempo y de calificarla con la clave que se les proporciona. Un cuadernillo similar, *Taking the SAT II: Subject Tests* ("La presentación del SAT II: Pruebas temáticas"), ilustra y explica los reactivos de diferentes pruebas temáticas.

Los Exámenes de Registro de Graduados (*Graduate Record Examinations*, GRE) también proporcionan materiales para familiarizarse con las pruebas. El *Information Bulletin* distribuido a todos los solicitantes comprende la explicación de una muestra de reactivos de la Prueba General (*General Test*), así como una prueba completa previamente aplicada con su clave de calificación. Para presentarla, se publican en un libro (*Practicing to Take the GRE General Test*) formas adicionales y también se dispone de cuadernillos prácticos similares que presentan pruebas individuales del GRE sobre diversas áreas temáticas.

El resurgimiento de los materiales de familiarización aparecidos en los ochenta y los noventa no se limita a los medios impresos, sino que incluye transparencias, diapositivas, películas, videocasetes y *software* para computadora. El Servicio de Pruebas Educativas (*Educational Testing Service*) ha realizado y distribuido muchos de estos materiales, y diseñó algunos para usarse con pruebas específicas, como es el caso de las diapositivas que acompañan a *Taking the SAT* y otros sobre la interpretación de los resultados del SAT y sobre las pruebas de logros de la Junta Universitaria. También se dispone de un programa de computadora para ayudar a comprender las puntuaciones

del SAT, y se ha elaborado un paquete de *software* para los estudiantes que quieren presentar la Prueba General del GRE. Por medio de un programa interactivo, el paquete contiene reactivos de muestra, una situación simulada de supervisión del tiempo, explicaciones de las preguntas contestadas de manera incorrecta y un análisis de los puntos fuertes y débiles del examinado.

Otros materiales (impresos, paquetes de multimedios, *software* para computadora) fueron diseñados para una orientación más general, y cubren temas que van de niños de escuela primaria a adultos. Un ejemplo es *On Your Own: Preparing for a Standardized Test* (1987), un videodisco para uso individual o grupal de estudiantes de secundaria. Una guía sencilla y completa en forma de libro es *How to Take a Test: Doing Your Best* (Dobbin, 1984). También las editoriales comerciales y algunas dependencias gubernamentales de los Estados Unidos han preparado guías para presentar las pruebas, como, por ejemplo, el conjunto de materiales para ser usados con la Batería de Pruebas de Aptitudes Generales (*General Aptitude Test Battery*, GATB) publicado por el Servicio de Empleo de los Estados Unidos.

Instrucción en habilidades cognoscitivas generales. Algunos investigadores han explorado el planteamiento opuesto a la mejora en el desempeño en la prueba. Su meta es la adquisición de habilidades intelectuales de gran aplicación, hábitos de trabajo y estrategias para la resolución de problemas. Los efectos de tales intervenciones deberían manifestarse *lo mismo* en los resultados de las pruebas que en el desempeño de criterio, como los cursos universitarios. De acuerdo con la distinción que presentamos al inicio de la sección, este programa está destinado a brindar educación más que preparación. Algunos de los investigadores que se ocupan del campo han estado trabajando con niños y adolescentes retardados educables (Babad y Budoff, 1974; Belmont y Butterfield, 1977; A. L. Brown, 1974; Budoff y Corman, 1974; Campione y Brown, 1979, 1987; Feuerstein, 1979, 1980; Feuerstein, Rand, Jensen, Kaniel y Tzuriel, 1987), mientras que otros han concentrado sus esfuerzos en estudiantes de educación media y superior con antecedentes de desventajas educativas (Linden y Whimbey, 1990; Whimbey, 1975, 1977, 1980).

Muchos de los procedimientos de capacitación empleados en esos programas fueron diseñados para desarrollar una conducta eficaz de solución de problemas: el análisis cuidadoso de los problemas o bien las preguntas, la consideración de todas las alternativas, los detalles pertinentes y las implicaciones de llegar a una solución, la formulación o elección de una solución deliberada más que impulsiva y la aplicación de criterios elevados para evaluar el propio desempeño. Se trata de estrategias que obviamente deberían mejorar el funcionamiento intelectual no sólo en las pruebas, sino también en el trabajo académico y en muchas otras actividades cotidianas que dependen del aprendizaje escolar; sin embargo, hay una pregunta crucial que tiene que ver con el grado de transferencia y generalización de los efectos aparte de los contenidos y los medios utilizados en la capacitación. Los resultados hasta ahora son prometedores, pero los programas aún se encuentran en etapa de exploración y se requiere de más investigaciones para establecer la amplitud y durabilidad de las mejoras alcanzadas.

Recapitulación. Hemos considerado tres formas de capacitación para las pruebas cuyos objetivos difieren considerablemente. ¿Cómo influyen en la validez de las prue-

bas y en su utilidad práctica como instrumentos de evaluación? La primera forma de entrenamiento es la preparación, en el sentido de un ejercicio intenso y masivo con reactivos similares a los de la prueba. Vimos que las pruebas bien elaboradas eligen aquellos que sean menos susceptibles a dicho ejercicio y los protegen. En tanto que semejante preparación pueda mejorar el desempeño en la prueba, lo hará sin mejorar la conducta de criterio, por lo que la validez del instrumento se reduce y se convierte en una medida menos eficaz de las habilidades generales que pretende evaluar y una forma menos precisa de indagar si el individuo ha adquirido las habilidades y los conocimientos que se requieren para tener éxito en la situación de criterio.

Por otro lado, los procedimientos de orientación sobre la prueba están diseñados para descartar o igualar las diferencias en las experiencias previas a su presentación. Del mismo modo que los efectos de la preparación, estas diferencias representan condiciones que influyen en los resultados de la prueba sin reflejarse necesariamente en el área general de conducta que pretende evaluarse, de ahí que los procedimientos de orientación aumenten la validez de los instrumentos al reducir la influencia de los factores relacionados con las pruebas.

Por último, la preparación en habilidades cognoscitivas de gran aplicación, cuando es eficaz, debe mejorar la habilidad del individuo para enfrentar las tareas intelectuales. Este progreso puede y debe reflejarse en los resultados de la prueba. En la medida en que mejoren tanto los resultados de la prueba como el desempeño en el criterio, esta preparación no modificará la validez de las pruebas y sí aumentará las posibilidades del individuo de alcanzar las metas deseadas.

FUENTES DE INFORMACIÓN

Las pruebas psicológicas se encuentran en un estado de rápido cambio. Las orientaciones se desplazan, hay una corriente constante de nuevas pruebas, de formas revisadas de pruebas antiguas y datos nuevos que pueden refinar o alterar la interpretación de los resultados. El ritmo acelerado de cambio, aunado al vasto número de pruebas existentes, hace que resulte poco práctico tratar de revisar pruebas concretas en un solo texto. En los libros que tratan del uso de los instrumentos en campos como la consejería, la práctica clínica, la selección de personal y la educación es posible encontrar una cobertura más amplia de los instrumentos y los problemas que enfrentan áreas especiales. En los capítulos respectivos anotamos las referencias a estas publicaciones.

De cualquier forma, todo el que trabaje con instrumentos de medición psicológicos necesita familiarizarse con las fuentes de información más directas para mantenerse al corriente. Una de las más importantes es el *Mental Measurements Yearbook* ("Anuario de medición mental") o MMY, establecido y editado por Oscar K. Buros desde 1938 (a partir de 1985, lo publica el Instituto Buros de Medición Mental de la Universidad de Nebraska). Esta serie de anuarios cubre casi todas las pruebas psicológicas, educativas y vocacionales publicadas en inglés y que están comercialmente disponibles. La cobertura es en especial completa para las pruebas de lápiz y papel. Cada anuario incluye las pruebas publicadas durante cierto periodo, por lo que no sustituye a los anuarios anteriores, sino que los complementa. Las primeras publicaciones de esta serie eran simples bibliografías de tests, pero desde 1938 adoptó su forma

actual, que incluye reseñas críticas de las pruebas a cargo de uno o más expertos, así como una lista completa de las referencias publicadas correspondientes a cada prueba. También proporciona información general sobre la editorial, el precio, las formas y la edad de los sujetos para los que es adecuada. El plan actual consiste en publicar un nuevo MMY cada dos o tres años y un suplemento entre cada anuario.

Las entradas del MMY, junto con las reseñas críticas, ahora están disponibles electrónicamente por medio de *SilverPlatter* (véase el apéndice B). La base de datos comienza con las entradas al noveno MMY y se actualiza cada seis meses. Otra publicación del Instituto Buros es *Tests in Print*, ahora en su cuarto volumen (TIP-IV, 1994), editado por L. L. Murphy, Conoley e Impara. Esta publicación proporciona una cobertura acumulada de todas las pruebas que se publican en inglés, junto con la información real y listas de referencias. Cada edición sucesiva también sirve como índice de todos los MMY anteriores.

Otra fuente importante de información sobre las pruebas publicadas son las *Test Collection Bibliographies*, preparadas por el Servicio de Pruebas Educativas (*Educational Testing Service*, ETS), que ofrece una bibliografía actualizada de las pruebas disponibles en áreas específicas de contenido. La cobertura es amplia y comprende todo tipo de pruebas, así como los instrumentos diseñados para usos particulares y poblaciones especiales, como los que padecen discapacidades físicas. Cada entrada contiene información real que incluye datos sobre el autor, la fecha de publicación, la editorial, la población objetivo, el propósito de la prueba y cualquier subpuntaje o variable por medir. Es posible adquirir las bibliografías de pruebas para áreas particulares por un costo nominal en *Test Collection*, ETS (la dirección aparece en el apéndice B). Ésta es una de varias publicaciones del ETS que brinda información actualizada sobre las pruebas y su aplicación.

Además de las pruebas publicadas, también hay una enorme cantidad de ellas descritas o reproducidas en libros, publicaciones periódicas o informes no publicados. De interés especial para los investigadores, estas pruebas han sido examinadas en diversos compendios (por ejemplo, Goldman y Mitchell, 1995). La información actualizada sobre las pruebas no publicadas se encuentra en *Tests in Microfiche*, distribuida por Test Collections, ETS. Cada año se agrega un nuevo conjunto de pruebas y se puede conseguir un índice de cada uno. Los usuarios calificados pueden adquirir pruebas aisladas o conjuntos. El directorio científico de la Asociación Psicológica Estadounidense (*Finding Information*, 1995), una fuente que se actualiza regularmente, tiene una guía clara y concisa para encontrar información sobre pruebas publicadas y no publicadas, y cualquiera que solicite una copia recibe automáticamente la versión más reciente.

Los usuarios encuentran la fuente más directa de información acerca de pruebas particulares en los catálogos de las editoriales y en el manual que acompaña a cada una. El *Mental Measurement Yearbook* contiene una lista completa de las editoriales especializadas y sus direcciones. Para una fácil referencia, en el apéndice B presentamos los nombres y domicilios de las editoriales cuyas pruebas citamos aquí. Se les puede solicitar los catálogos de pruebas actuales, pero los manuales y las pruebas sólo están a disposición de los usuarios calificados.

El manual de la prueba debe ofrecer la información esencial que se necesita para aplicarla, calificarla y evaluarla; debe incluir instrucciones completas y detalladas, la

clave de calificación, las normas y los datos sobre confiabilidad y validez, y, además, debe informar del número y la naturaleza de las personas en las que se establecieron las normas, la confiabilidad y la validez, así como los métodos utilizados para calcular los índices de estas medidas. En caso de que la información necesaria sea demasiado extensa para el manual, debe dar las referencias al manual técnico o a otros medios impresos en los que se encuentre. En otras palabras, el manual permite que el usuario evalúe la prueba antes de elegirla para sus propósitos particulares. Agreguemos que algunos manuales aún están lejos de esta meta, pero las grandes y más profesionales editoriales especializadas brindan cada vez mayor atención a la preparación de manuales que cumplan los criterios científicos adecuados. Es de esperar que un público ilustrado de usuarios sea la mejor garantía de que dichos criterios se mantendrán y mejorarán.

Es posible encontrar una guía completa, aunque sucinta, para la evaluación de las pruebas psicológicas en los *Standards for Educational and Psychological Testing* ("Criterios para las Pruebas Educativas y Psicológicas") preparados por la Asociación Estadounidense de Psicología (*American Psychological Association*, APA) en colaboración con otras dos asociaciones interesadas en las pruebas, la Asociación Estadounidense de Investigación Educativa (*American Educational Research Association*, AERA) y el Consejo Nacional de Medición Educativa (*National Council on Measurement in Education*, NCME). Publicados inicialmente en 1954, los "Criterios" han sido revisados en 1966, 1974 y 1985; en la actualidad, está en marcha una nueva revisión por parte de las tres asociaciones participantes.

En los ochenta surgió la necesidad de establecer criterios para las pruebas (*Testing Standards*[2]) que no sólo se preocuparan por la calidad técnica de las pruebas sino también por su efecto sobre el bienestar del individuo (véase la figura 1.1, página 30). La naturaleza de la revisión más reciente de los *Testing Standards* indica que esta preocupación es una tendencia en progreso. La figura 1.2 de la página 31 contiene una lista propuesta de criterios preparada por una comisión de las tres asociaciones en 1966. Es evidente que el interés por adaptar la selección de pruebas —así como la preocupación por la interpretación y el uso de sus resultados— al conocimiento sobre las experiencias del examinado muestra un crecimiento continuo. Es digno de observar que toda una sección (Segunda Parte) de la figura 1.2 se titula "Equidad de la prueba". Los usuarios han cobrado conciencia de que la aplicación inadecuada de los instrumentos puede dañar al individuo y disminuir la eficacia de sus contribuciones a la sociedad. Las críticas populares por el mal uso de las pruebas puede haber contribuido a esta mayor conciencia de los examinadores, que a su vez debe disminuir los abusos y, al mismo tiempo, aumentar el reconocimiento público de los beneficios del uso de las pruebas.

[2] Para abreviar, en adelante seguiremos la práctica común de identificarlos como Testing Standards.

CRITERIOS PARA LA APLICACIÓN DE PRUEBAS EDUCATIVAS Y PSICOLÓGICAS

Primera parte. Criterios técnicos para la elaboración y evaluación de los instrumentos

1. Validez
2. Confiabilidad y errores de medición
3. Desarrollo y revisión
4. Escalamiento, normatividad, comparabilidad de las calificaciones e igualdad
5. Publicación: manuales técnicos y guías para el usuario

Segunda parte. Criterios profesionales para el uso de los instrumentos

6. Principios generales de uso
7. Pruebas clínicas
8. Pruebas educativas y aplicación de pruebas psicológicas en las escuelas
9. Uso de pruebas en la consejería
10. Pruebas laborales
11. Licencia y certificación profesional y ocupacional
12. Programa de evaluación

Tercera parte. Criterios para aplicaciones particulares

13. Evaluación de minorías lingüísticas
14. Evaluación de gente con condiciones discapacitantes

Cuarta parte. Criterios para los procedimientos administrativos

15. Aplicación, calificación y reporte
16. Protección de los derechos de los evaluados

Figura 1.1. Temas cubiertos por los *Standards for Educational and Psychological Testing* (AERA, APA, NCME, 1985).

LISTA PROPUESTA DE CRITERIOS PARA LOS INSTRUMENTOS EDUCATIVOS Y PSICOLOGICOS

Primera parte. Elaboración, evaluación y documentación de las pruebas

1. Validez
2. Confiabilidad, errores de medición y función informativa de las calificaciones
3. Desarrollo y revisión del instrumento
4. Escalamiento, normas, criterios y comparación de las calificaciones
5. Aplicación, calificación y reporte
6. Documentos de la prueba

Segunda parte. Equidad de las pruebas

7. Equidad y sesgos
8. Protección de los derechos del examinado
9. Aplicación de las pruebas a personas cuya primera lengua no es el inglés
10. Aplicación de las pruebas a individuos con discapacidades

Tercera parte. Aplicaciones de las pruebas

11. Principios generales de uso
12. Pruebas para la evaluación psicológica
13. Pruebas para la evaluación educativa
14. Pruebas laborales, de licencia y certificación
15. Pruebas para los programas de evaluación y política pública

Figura 1.2. Temas elegidos para la edición revisada de los *Standars for Educational and Psychological Testing* (AERA, APA, NCME, 1966). Manuscrito en preparación.

(Reproducido con autorización del Comité Conjunto para los *Standards for Educational and Psychological Testing* [Dianne Brown, directora del proyecto]).

Capítulo Dos

Antecedentes históricos de las pruebas actuales

Una breve revisión de los antecedentes históricos y los orígenes de las pruebas psicológicas nos brindará un panorama útil para comprender las pruebas actuales.[1] La dirección en que éstas han avanzado, así como sus limitaciones y ventajas especiales, podrán entenderse mejor al considerarlas a la luz de sus precursores. En este capítulo nos concretaremos a la revisión de los antecedentes y el desarrollo inicial del movimiento psicométrico como un todo. En los capítulos posteriores analizaremos los desarrollos más recientes relacionados con algunos instrumentos específicos, como los tests de habilidad (capítulos 8 a 12) o los de interés (capítulo 14), así como con las áreas (educativa, industrial, clínica y de consejería) en las que se utilizan tales instrumentos (capítulo 17).

Las raíces de la aplicación de pruebas se pierden en la Antigüedad. Existen relatos del sistema de exámenes del servicio civil que prevaleció en el imperio chino durante 2 000 años (Bowman, 1989). Entre los antiguos griegos, la aplicación de exámenes formaba parte del proceso educativo; las pruebas servían para evaluar el dominio de habilidades físicas e intelectuales (Doyle, 1974). Desde sus inicios en la Edad Media, las universidades europeas basaron los grados y honores en exámenes formales. Con todo, no tenemos que ir más allá del siglo XIX para identificar los

[1] Se encuentra una descripción más detallada de los orígenes de las pruebas psicológicas en F. L. Goodenough (1949) y en J. Peterson (1926). Véase también Boring (1950) y G. Murphy y Kovach (1972) para antecedentes más generales, DuBois (1970) y McReynolds (1975, 1986) para recuentos más recientes de la historia de las pruebas psicológicas, y Anastasi (1965) para los antecedentes históricos del estudio de las diferencias individuales. En Anastasi (1993) hay un repaso general de las tendencias actuales de la psicometría.

principales acontecimientos que forman la base de las pruebas contemporáneas, y a ellos dirigimos ahora nuestra atención.

INTERÉS INICIAL EN LA CLASIFICACIÓN Y LA CAPACITACIÓN DE LAS PERSONAS CON RETARDO MENTAL

El siglo XIX atestiguó el surgimiento del interés por el tratamiento humano de las personas "insanas" y las que sufrían de retardo mental. Hasta ese momento, el destino común de tales individuos era el descuido, el ridículo e incluso la tortura. Con la nueva preocupación por el cuidado adecuado de la gente con problemas mentales, se hizo evidente la necesidad de contar con criterios uniformes para su identificación y clasificación, necesidad que se volvió verdaderamente urgente con la proliferación de instituciones sociales dedicadas a estas personas en todo el mundo. Primero era necesario distinguir entre los individuos insanos y los que sufrían de retardo mental. Los primeros manifestaban trastornos emocionales que podrían o no estar acompañados por un deterioro intelectual a partir de un nivel normal; los segundos se caracterizaban principalmente por una deficiencia intelectual que estaba presente desde el nacimiento o la primera infancia. El primer informe explícito de esta distinción se encuentra quizá en un trabajo publicado en 1838 por el médico francés Esquirol. Se trata de dos volúmenes en los que dedica más de 100 páginas a lo que ahora se conoce como "retardo mental". Esquirol también señala que existen muchos grados de retardo, que varían desde la normalidad hasta la "idiotez profunda". En su esfuerzo por elaborar un sistema para clasificar los diversos grados y variedades de retardo, Esquirol probó varios procedimientos y llegó a la conclusión de que el habla proporciona el criterio más confiable del nivel intelectual del individuo. Es importante decir que los criterios actuales para determinar el retardo mental también son principalmente lingüísticos y que los tests actuales de inteligencia tienen una fuerte carga de contenido verbal. En los capítulos siguientes mostraremos la importante función que cumple la habilidad verbal en nuestro concepto de inteligencia.

De especial significado son las contribuciones de otro médico francés, Seguin, quien fue pionero en la capacitación de los retardados. Luego de rechazar la idea dominante de que el retardo mental era incurable, Seguin (1866/1907) experimentó durante muchos años con lo que llamó el método fisiológico de capacitación, y en 1837 estableció la primera escuela dedicada a la educación de los niños retardados. En 1848 emigró a los Estados Unidos, en donde sus ideas obtuvieron un amplio reconocimiento. Seguin fue el creador de muchas de las técnicas de entrenamiento sensorial y muscular que después fueron adoptadas por las instituciones de asistencia a los retrasados. Con esos métodos, los niños con retardo profundo reciben ejercicio intensivo en discriminación sensorial y para el desarrollo del control motor. Algunos de los procedimientos que estableció Seguin con este propósito fueron luego incorporados a los tests de inteligencia no verbal o de ejecución. Un ejemplo de lo anterior es el

Tablero de Formas* de Seguin (*Seguin Form Board*), en el que el individuo debe insertar bloques de distinta apariencia en los huecos correspondientes tan rápidamente como pueda.

Más de medio siglo después del trabajo de Esquirol y Seguin, el psicólogo francés Alfred Binet recomendaba que los niños que no lograran responder a la educación normal fueran evaluados antes de expulsarlos y que, de ser considerados educables, fueran asignados a cursos de educación especial (T. H. Wolf, 1973). Con sus compañeros de la Sociedad para el Estudio Psicológico del Niño, Binet exhortó al ministro de Instrucción Pública a emprender acciones para mejorar las condiciones de los niños retardados. Un resultado especial fue el establecimiento de una comisión ministerial a cargo de Binet para el estudio de estos infantes. Tal designación fue un acontecimiento trascendental en la historia de las pruebas psicológicas.

LOS PRIMEROS PSICÓLOGOS EXPERIMENTALES

En general, los primeros psicólogos experimentales del siglo XIX no estaban interesados en la medición de las diferencias individuales. Su objetivo principal era la formulación de descripciones generalizadas de la conducta humana, lo que supone que su atención se concentraba en las uniformidades más que en las diferencias conductuales, y que, por lo tanto, las diferencias individuales eran ignoradas o aceptadas como un mal necesario que limitaba la aplicabilidad de las generalizaciones. Así, el hecho de que un individuo reaccionara de manera diferente a otro en las mismas condiciones era considerado como un error, o variabilidad individual, que disminuía la exactitud de las generalizaciones. Ésta era la actitud hacia las diferencias individuales que prevalecía en laboratorios como el fundado por Wundt en Leipzig en 1879, donde se formaron muchos de los primeros psicólogos experimentales.

Al elegir sus áreas de estudio, como en muchas otras fases de su trabajo, los fundadores de la psicología experimental evidenciaban la influencia de sus antecedentes en la fisiología y la física. Los problemas que estudiaban en sus laboratorios tenían que ver principalmente con el tiempo de reacción y con la sensibilidad a estímulos sensoriales como los visuales y los auditivos. Como veremos en otros capítulos, este acento en los fenómenos sensoriales refleja también la naturaleza de las primeras pruebas psicológicas.

Otra influencia que la psicología experimental del siglo XIX ejerció en el curso del movimiento psicométrico se manifiesta en su insistencia por ejercer un control riguroso de las condiciones en las que se realizan las observaciones; por ejemplo, la forma de dar las instrucciones en un experimento de tiempo de reacción puede aumentar o disminuir de modo notable la velocidad de la respuesta, y el brillo o el color del fondo pueden alterar marcadamente la apariencia de un estímulo visual, lo que demuestra la importancia de que todos los participantes del experimento sean observados en condiciones estandarizadas. Esta estandarización del procedimiento llegó a convertirse en una de las señales distintivas de las pruebas psicológicas.

*También conocido como *Tablero de encajamientos*. (N. del T.)

CONTRIBUCIONES DE FRANCIS GALTON

El biólogo inglés Francis Galton fue el principal responsable del inicio del movimiento psicométrico. Un factor común en las numerosas y variadas actividades de investigación de Galton fue su interés por la herencia humana. En el curso de sus investigaciones sobre esta materia, Galton se dio cuenta de la necesidad de medir las características de personas que estaban y no emparentadas, ya que sólo de esta manera podía descubrir, por ejemplo, el grado exacto de parecido entre padres e hijos, hermanos y hermanas, primos o gemelos. Con este propósito, Galton convenció a diversas instituciones educativas de que llevaran registros antropométricos sistemáticos de sus estudiantes. También estableció un laboratorio antropométrico en la Exposición Internacional de 1884 donde, mediante el pago de tres peniques, se medían ciertos rasgos físicos de los visitantes y se los sometía a pruebas de agudeza visual y auditiva, fuerza muscular, tiempo de reacción y otras funciones sensoriomotoras simples. Cuando la exposición cerró, Galton transfirió el laboratorio al Museo de South Kensington, en Londres, donde trabajó durante seis años. Con estos métodos se acumuló gradualmente el primer cuerpo sistemático de datos sobre diferencias individuales en los procesos psicológicos simples.

El propio Galton elaboró varias de las pruebas aplicadas en su laboratorio, muchas de las cuales siguen siendo conocidas, sea en su forma original o modificada. Entre los ejemplos se encuentran la barra de Galton para la discriminación visual de la longitud, el silbato que lleva también su nombre para determinar el mayor tono audible y una serie graduada de pesos para medir la discriminación cinestésica. Galton creía que las pruebas de discriminación sensorial podían servir para estimar el intelecto. A este respecto, había sido también la influencia de las teorías de Locke. Así, Galton escribió: "La única información concerniente a los acontecimientos externos que nos alcanza parece pasar por la avenida de los sentidos, y entre más perceptivos sean los sentidos de las diferencias, mayor será el campo sobre el que pueden actuar nuestro juicio y nuestra inteligencia" (Galton, 1883, p. 27). Galton también se percató de que las personas con retardo mental extremo tienden a mostrar defectos en su capacidad para discriminar entre calor, frío y dolor, una observación que posteriormente fortaleció su convicción de que la capacidad de discriminación sensorial, "en conjunto, sería más elevada entre los intelectualmente aptos" (Galton, 1883, p. 29).

Galton también fue pionero en la aplicación de escalas de calificación y cuestionarios, así como en el uso de la técnica de asociación libre que posteriormente fue utilizada con diversos propósitos. Otra de sus muchas contribuciones se encuentra en la formulación de métodos estadísticos para el análisis de datos sobre las diferencias individuales. Galton tomó y adaptó algunas técnicas matemáticas de forma que pudieran ser utilizadas por investigadores sin conocimientos en la materia que desearan tratar en forma cuantitativa los resultados de sus pruebas. De esta manera extendió considerablemente la aplicación de los procedimientos estadísticos al análisis de datos de las pruebas. Esta fase del trabajo de Galton fue llevada adelante por muchos de sus estudiantes, de los cuales Karl Pearson[2] fue el más eminente.

[2] Para una descripción fascinante de los primeros avances de los principales conceptos estadísticos y de las personas responsables, véase Cowles (1989).

CATTELL Y LOS PRIMEROS TESTS MENTALES

El psicólogo estadounidense James McKeen Cattell ocupa una posición prominente en el adelanto de las pruebas psicológicas. En su trabajo se combinan la recién establecida ciencia de la psicología experimental y el todavía más reciente movimiento psicométrico. Para obtener su doctorado en Leipzig redactó una tesis sobre el tiempo de reacción bajo la dirección de Wundt; y en 1888, mientras enseñaba en Cambridge, su trato con Galton fortaleció su interés por la medición de las diferencias individuales. A su regreso a los Estados Unidos, Cattell participó activamente en el establecimiento de laboratorios de psicología experimental y en la difusión del movimiento psicométrico.

El término "test mental" apareció por primera vez en la bibliografía psicológica en un artículo escrito por Cattell en 1890, que describe una serie de pruebas que cada año se aplicaban a los estudiantes universitarios para determinar su nivel intelectual. Las pruebas, cuya aplicación era individual, incluían mediciones de la fuerza muscular, velocidad de movimiento, sensibilidad al dolor, agudeza visual y auditiva, discriminación de pesos, tiempo de reacción, memoria y cosas similares. En su elección de las pruebas, Cattell compartía la opinión de Galton respecto a la posibilidad de obtener una medida de las funciones intelectuales con el uso de instrumentos de discriminación sensorial y de tiempo de reacción. La preferencia de Cattell por dichas herramientas también se sustentó en el hecho de que le permitían medir con precisión y exactitud las funciones simples, mientras que la obtención de mediciones objetivas de funciones más complejas en ese tiempo parecía una tarea inalcanzable.

Las pruebas de Cattell son características de los instrumentos desarrollados en la última década del siglo XIX, que se aplicaban a escolares, universitarios y adultos. En la Exposición de Columbia realizada en Chicago durante 1893, Jastrow montó una exhibición en la que a los visitantes se los invitaba a realizar pruebas de procesos perceptuales, sensoriales y motores simples y a comparar su habilidad con las normas (J. Peterson, 1926, Philippe, 1894). Los pocos intentos por evaluar estas primeras pruebas arrojaron resultados desalentadores. El desempeño de cada individuo mostraba poca correspondencia de una prueba a otra (Sharp, 1898–1899; Wissler, 1901) y su relación con estimaciones independientes del nivel intelectual basadas en las calificaciones de los maestros (T. L. Bolton, 1891–1892; J. A. Gilbert, 1894) o con las calificaciones académicas (Wissler, 1901) era poca o nula.

Algunas pruebas elaboradas por psicólogos europeos de la época tendían a cubrir funciones algo más complejas. Kraepelin (1895), que estaba interesado sobre todo en la evaluación clínica de pacientes psiquiátricos, preparó una larga serie de pruebas para medir lo que consideraba factores básicos en la caracterización del individuo. Las pruebas, que empleaban principalmente operaciones aritméticas simples, estaban destinadas a medir los efectos de la práctica, la memoria y la susceptibilidad a la fatiga y la distracción. Otro psicólogo alemán, Ebbinghaus (1897), aplicó a escolares pruebas de cálculo aritmético, memoria y completación de oraciones; esta última, que era la más compleja, fue la única que mostró una clara correspondencia con el desempeño escolar.

En un artículo publicado en Francia en 1895, Binet y Henri criticaron el hecho de que casi todas las pruebas disponibles eran sensoriales y se concentraban indebidamente en habilidades especializadas simples. Además, argumentaban que en la medición de las funciones más complejas no se requiere de gran precisión, ya que, en tales funciones, las diferencias individuales son mayores. Propusieron una amplia y variada lista de tests que cubrían funciones como la memoria, la imaginación, la atención, la comprensión, la susceptibilidad a la sugestión, la apreciación estética y muchos otros, en los que podemos reconocer las tendencias que a la postre condujeron al desarrollo de las famosas escalas de inteligencia de Binet.

BINET Y EL SURGIMIENTO DE LOS TESTS DE INTELIGENCIA

Binet y sus colaboradores dedicaron muchos años a la investigación activa e ingeniosa de las formas de medir la inteligencia. Probaron muchos métodos, incluyendo la medición de la forma del cráneo, la cara y la mano, así como el análisis de la escritura; sin embargo, los resultados los llevaron a la convicción de que la medición directa, aunque tosca, de las funciones intelectuales complejas era la más promisoria. Entonces una situación en particular hizo fructificar los esfuerzos de Binet. En 1904, el Ministerio de Educación lo comisionó para que estudiara procedimientos para la educación de niños retardados. Fue en relación con los objetivos de esta comisión que Binet preparó, en colaboración con Simon, la primera escala de Binet-Simon (Binet y Simon, 1905).

Esta escala, conocida como la escala de 1905, constaba de 30 problemas o tests arreglados en orden de dificultad creciente. El nivel de dificultad se estableció empíricamente aplicándolos a 50 niños normales de tres a 11 años y a algunos niños y adultos retardados. Las pruebas fueron diseñadas para cubrir una amplia variedad de funciones, con énfasis especial en el juicio, la comprensión y el razonamiento, que Binet consideraba los componentes principales de la inteligencia. Aunque incluyeron pruebas sensoriales y perceptuales, en esta escala se encuentra una proporción de contenido verbal mayor que en la generalidad de las pruebas de la época. Como la escala de 1905 fue presentada como un instrumento preliminar y tentativo, no se formuló un método objetivo preciso para obtener una puntuación total.

En la segunda escala, la de 1908, aumentó el número de tests, se eliminaron algunos de la primera que resultaron insatisfactorios y todos fueron agrupados en niveles de edad sobre la base del desempeño de alrededor de 300 niños normales de entre tres y 13 años. De este modo, en el nivel de tres años se ubicaron todos los tests que pasaban del 80 al 90 por ciento de los niños normales de tres años; en el nivel de cuatro años, los que aprobó el mismo porcentaje de niños normales de esa edad; y así sucesivamente hasta los 13. La calificación del niño en toda la prueba podía entonces expresarse como el *nivel mental* correspondiente a la edad de los niños normales cuya ejecución había igualado. En las diversas traducciones y adaptaciones de las escalas de Binet el término de "nivel mental" fue sustituido por el de "edad mental", cuya fácil comprensión indudablemente contribuyó a popularizar los tests de

inteligencia;[3] sin embargo, el propio Binet evitaba el uso del término por sus implicaciones no verificadas de desarrollo y prefería el más neutral de "nivel mental" (T. H. Wolf, 1973).

En 1911, año en que Binet murió intempestivamente, apareció la tercera revisión de la Escala de Binet-Simon, que no presenta mayores modificaciones, salvo revisiones menores, cambios en la localización de algunos tests, la adición de otros en varios niveles de edad y la extensión de la escala al nivel adulto.

Incluso antes de la revisión de 1908, los tests de Binet-Simon atrajeron la atención de los psicólogos de todo el mundo. En muchos países, aparecieron traducciones y adaptaciones, pero la primera fue la de H. H. Goddard, en ese entonces psicólogo investigador en la Escuela de Capacitación de Vineland para niños con retardo mental. La revisión de Goddard resultó clave para que la profesión médica aceptara los tests de inteligencia (Zenderland, 1987). Apareció en un momento propicio para satisfacer la urgente necesidad de un instrumento estandarizado para diagnosticar y clasificar a las personas con retardo mental. No obstante, como herramienta de evaluación pronto fue dejada atrás por el Stanford-Binet, un instrumento más extenso y psicométricamente refinado, elaborado por L. M. Terman y sus colaboradores en la Universidad de Stanford (Terman, 1916). Este instrumento utilizó por primera vez el cociente de inteligencia (CI) o razón entre la edad mental y la cronológica. La última revisión se emplea extensamente y la estudiaremos de manera más completa en el capítulo 8. También resulta de especial interés la primera revisión Kuhlmann-Binet, que amplió la escala hasta la edad de tres meses (Kuhlmann, 1912) y representa uno de los primeros esfuerzos por elaborar tests de inteligencia para preescolares e infantes.

PRUEBAS COLECTIVAS

Los tests de Binet, así como todas sus revisiones, son *escalas individuales* en el sentido de que sólo pueden aplicarse a una persona y luego a otra. Muchos de los tests de esas escalas requieren respuestas orales del examinado o la manipulación de materiales. Algunos exigen tomar el tiempo de las respuestas de cada individuo. Por éstas y otras razones, las pruebas no pueden adaptarse a la aplicación colectiva. Otra característica de los tests de Binet es que requieren de un examinador muy capacitado, pues se trata de instrumentos esencialmente clínicos adecuados para el estudio intensivo de casos individuales.

Las *pruebas de grupo* similares a la primera escala de Binet fueron elaboradas para satisfacer una necesidad práctica. Cuando los Estados Unidos ingresaron en 1917 en la Primera Guerra Mundial, la Asociación Estadounidense de Psicología formó una comisión para considerar cómo podía contribuir la psicología a la conducción de la

[3] F. L. Goodenough (1949, pp. 50–51) observó que en 1887, 21 años antes de la aparición de la Escala Binet-Simon de 1908, S. E. Chaille publicó en el *New Orleans Medical and Surgical Journal* una serie de pruebas para infantes, arregladas de acuerdo con la edad en que solían ser pasadas. Debido en parte a la limitada circulación de la revista y quizá también a que la comunidad científica no estaba preparada, en esa época pasó inadvertido el significado del concepto de escala de edad. La propia escala de Binet recibió la influencia del trabajo de algunos de sus contemporáneos, principalmente Blin y Damaye, que prepararon un conjunto de preguntas orales de las que obtenían una sola calificación global para cada niño (T. H. Wolf, 1973).

guerra. La comisión, dirigida por Robert M. Yerkes, reconoció la necesidad de la rápida clasificación del nivel intelectual general del millón y medio de reclutas. Esta información era importante para tomar muchas decisiones administrativas, incluyendo el rechazo o la dispensa del servicio militar, la asignación a diferentes servicios o la admisión a campos de entrenamiento de oficiales. En este ambiente se formuló el primer test colectivo de inteligencia. Para realizar la tarea, los psicólogos acudieron a todos los materiales disponibles, especialmente a un test colectivo de inteligencia no publicado, preparado por Arthur S. Otis, que cedió al ejército. Una contribución importante de este instrumento, elaborado por Otis cuando era estudiante en uno de los cursos de grado de Terman, fue la introducción de preguntas de opción múltiple y otros reactivos "objetivos".

Las pruebas que finalmente desarrollaron los psicólogos del ejército podían aplicarse a grupos grandes y llegaron a conocerse como Army Alpha y Army Beta (pruebas Alfa y Beta). La primera se ocupaba de las pruebas generales de rutina, mientras que la segunda era una escala no lingüística utilizada con reclutas iletrados o procedentes de países cuyo idioma no era el inglés.

Poco después de terminar la Primera Guerra Mundial, las pruebas del ejército fueron liberadas para que se aplicaran a civiles; además de pasar por muchas revisiones, las pruebas Alfa y Beta sirvieron como modelo para la mayor parte de los tests colectivos de inteligencia, lo que contribuyó al notable crecimiento del movimiento psicométrico. Muy pronto se prepararon tests colectivos de inteligencia para todas las edades y personas, desde preescolares hasta universitarios. Los programas de evaluación a gran escala, otrora imposibles, se emprendieron con gran entusiasmo. Como las pruebas colectivas fueron diseñadas como instrumentos de evaluación masiva, no sólo permitieron la medición simultánea de grandes grupos, sino que también simplificaron las instrucciones y los procedimientos de aplicación, con lo que disminuyó la capacitación requerida del aplicador. Los maestros empezaron a aplicar en sus grupos los tests de inteligencia y los estudiantes universitarios eran evaluados de rutina antes de su ingreso; se iniciaron amplios estudios de grupos especiales de adultos, como los prisioneros, y muy pronto el público general cobró conciencia del CI.

La aplicación de esos tests colectivos excedió a su mejoramiento técnico. En la prisa por obtener puntuaciones y sacar conclusiones prácticas de los resultados, a menudo se olvidaba que las pruebas aún eran técnicamente imperfectas, de modo que cuando no lograban cumplir las injustificadas expectativas, se generaba escepticismo y hostilidad hacia todas las pruebas. De este modo, el auge psicométrico de los años veinte, basado en el uso indiscriminado de los instrumentos, puede haber demorado el progreso de las pruebas psicológicas.

TESTS DE APTITUD

Aunque al principio los tests de inteligencia fueron diseñados para cubrir una amplia variedad de funciones que permitieran estimar el nivel general de inteligencia del individuo, pronto se hizo evidente que su alcance era muy limitado. No todas las funciones importantes estaban representadas. De hecho, la mayor parte de los tests de inteligencia eran principalmente mediciones de la habilidad verbal y, en menor grado, de la

habilidad para manejar relaciones numéricas y otras relaciones abstractas y simbólicas. Gradualmente, los psicólogos reconocieron que el término "test de inteligencia" era un nombre inadecuado, ya que sólo medían ciertos aspectos de ésta.

Para que resultaran seguros, los instrumentos tocaban habilidades que son de gran importancia en la cultura para la que fueron diseñados, pero pronto se reconoció la conveniencia de contar con designaciones más precisas en términos de la información que podían ofrecer; por ejemplo, ahora se conoce como tests de aptitud escolar a algunos instrumentos que en los veinte habrían sido llamados de inteligencia. Este cambio en la terminología procede del hecho de que muchos de los llamados tests de inteligencia miden la combinación de habilidades requeridas y fomentadas por el trabajo universitario.

Incluso antes de la Primera Guerra Mundial, los psicólogos empezaban a admitir la necesidad de contar con tests de aptitudes especiales que complementaran los de inteligencia global. En especial, se prepararon *pruebas de aptitudes especiales* para uso en la orientación vocacional y en la selección y clasificación de personal industrial y militar. Entre las más empleadas están los tests de aptitud mecánica, profesional, musical y artística.

La evaluación crítica de las pruebas de inteligencia que siguió al uso indiscriminado durante los veinte también reveló el hecho notable de que el desempeño del individuo a menudo mostraba una marcada variación en diferentes partes de la prueba. Esto resultó en especial evidente en las pruebas colectivas, en las que los reactivos suelen separarse en subpruebas de contenido relativamente homogéneo; por ejemplo, una persona podía obtener una puntuación más bien alta en la subprueba verbal y un pobre resultado en la numérica o viceversa. Dicha variabilidad interna es hasta cierto grado discernible en un test como el Stanford-Binet en el que, por ejemplo, todos los reactivos que incluyen palabras pueden resultar difíciles para cierto individuo, mientras que los que emplean imágenes o diagramas geométricos pueden resultarle ventajosos.

Los usuarios, y sobre todo los clínicos, a menudo se valían de esas comparaciones internas para obtener un conocimiento más profundo de la estructura psicológica del individuo. De este modo, al evaluar un caso individual no sólo se examinaba el CI o algún otro resultado global, sino también el desempeño en grupos de reactivos o subtests; sin embargo, en general no se recomienda esta práctica, porque los tests de inteligencia no fueron diseñados con el propósito de realizar un análisis diferencial de las aptitudes. Con frecuencia, los subtests comparados contienen muy pocos reactivos para producir una estimación estable o confiable de determinada habilidad. Entonces, la diferencia de las puntuaciones de los subtests se puede revertir si vuelve a examinarse al individuo otro día o con otra forma del mismo instrumento. Para realizar esas comparaciones en el mismo sujeto, es necesario diseñar las pruebas con el objeto de que revelen diferencias de ejecución en diversas funciones.

Al mismo tiempo que la aplicación práctica demostraba la necesidad de contar con múltiples tests de aptitud, un desarrollo paralelo en el estudio de la organización de los rasgos iba proporcionando los medios para elaborarlos. Los estudios estadísticos sobre la naturaleza de la inteligencia habían explorado las relaciones entre las puntuaciones obtenidas por muchas personas en una gran variedad de pruebas. El psicólogo inglés

Charles Spearman (1904, 1927) inició estas investigaciones durante la primera década del siglo XX. Los desarrollos metodológicos subsecuentes, basados en el trabajo de estudiosos ingleses y estadounidenses, como T. L. Kelley (1928) y L. L. Thurstone (1938, 1947b), llegaron a conocerse como *análisis factorial*.

En el capítulo 11 examinaremos de manera exhaustiva las contribuciones de los métodos del análisis factorial en la elaboración de pruebas. Por ahora basta con destacar que los datos obtenidos con estos procedimientos señalaron la presencia de diversos factores o rasgos relativamente independientes, algunos representados en diversas proporciones en los tests tradicionales de inteligencia, como, por ejemplo, la comprensión verbal y el razonamiento numérico; otros, como la aptitud espacial, perceptual y mecánica, se encuentran más a menudo en los instrumentos para la medición de aptitudes especiales que en los de inteligencia.

Uno de los principales resultados prácticos del análisis factorial fue el desarrollo de las *baterías de aptitudes múltiples*, diseñadas para proporcionar una medida de la posición del individuo en una serie de rasgos. En lugar de una puntuación total, o CI, estos instrumentos obtienen una puntuación aparte para rasgos como la comprensión verbal, la aptitud numérica, la visualización espacial, el razonamiento aritmético y la rapidez perceptual, lo que permite realizar el análisis intrasujeto o diagnóstico diferencial en el mismo sujeto que los usuarios trataron de obtener por muchos años con los resultados crudos y a menudo erróneos de los tests de inteligencia. Además, dado que cubren algunos de los rasgos que generalmente no se incluían en ese tipo de instrumentos también incorporaron en un programa amplio y sistemático de evaluación gran parte de la información que antes se obtenía de las pruebas de aptitud especial.

Las baterías de aptitudes múltiples representan un desarrollo relativamente tardío en el campo de las pruebas, ya que casi todas aparecieron a partir de 1945. A este respecto es de destacar el trabajo de los psicólogos militares durante la Segunda Guerra Mundial. Gran parte de la investigación conducida en las fuerzas armadas se basó en el análisis de factores y se orientó a la preparación de baterías de aptitudes múltiples; por ejemplo, en la fuerza aérea se construyeron baterías especiales para pilotos, bombarderos, operadores de radio, rastreadores y muchos otros especialistas. El informe de las baterías que elaboró la fuerza aérea ocupa al menos nueve de los 19 volúmenes dedicados al programa psicológico de la aviación durante la Segunda Guerra Mundial (*Army Air Forces*, 1947–1948). La investigación en esas líneas aún sigue en progreso con el patrocinio de varias ramas de las fuerzas armadas. También se han realizado baterías para uso civil y se aplican lo mismo en la orientación vocacional y educativa que en la selección y clasificación de personal. En los capítulos 10 y 17 presentaremos algunos ejemplos.

Un adelanto más reciente, surgido a finales de los ochenta y comienzos de los noventa, permite la integración de dos métodos al principio contradictorios de medición mental representados por los tests tradicionales de inteligencia y las baterías de aptitudes múltiples (Anastasi, 1994). En la actualidad, se reconoce que la habilidad del ser humano puede ser evaluada de manera adecuada a diferentes niveles de profundidad, desde las aptitudes muy definidas de las pruebas específicas (o incluso de reactivos), pasando por el nivel de los rasgos hasta una puntuación global como el tradicional CI. Para cada propósito de evaluación hay un nivel apropiado de profundidad. En

correspondencia, los tests de inteligencia que se han desarrollado recientemente, como las Escalas de Habilidad Diferencial o las revisiones recientes de tests anteriores, como la cuarta edición del Stanford-Binet (que explicaremos en el capítulo 8), combinan la amplia cobertura de diversas aptitudes con calificaciones flexibles de niveles múltiples para propósitos concretos de evaluación. Aunque se trata de dos ejemplos de tests de inteligencia de aplicación individual, el mismo método amplio y flexible de preparación y uso de los instrumentos ha tenido un impacto en las baterías de aplicación colectiva, como las que veremos en el capítulo 10. En el capítulo 11 estudiaremos la base teórica y las implicaciones prácticas de esta combinación de tests de habilidad en relación con los adelantos recientes, relativos a la naturaleza de la inteligencia.

PRUEBAS ESTANDARIZADAS DE APROVECHAMIENTO

Mientras los psicólogos se ocupaban de los tests de inteligencia y de aptitud, los exámenes escolares tradicionales experimentaban avances técnicos (O. W. Caldwell y Courtis, 1923; Ebel y Damrin, 1960). Un paso importante en esta dirección lo dieron las escuelas públicas de Boston en 1845 cuando los exámenes escritos sustituyeron al interrogatorio oral de los estudiantes por examinadores visitantes. Entre los argumentos ofrecidos en apoyo de esta innovación estaba que los exámenes escritos ponían a los estudiantes en una situación uniforme, permitían una cobertura más amplia del contenido, reducían el elemento azaroso en la elección de reactivos y eliminaban la posibilidad de favoritismo por parte del examinador. Todos estos argumentos tienen un sonido familiar: fueron utilizados mucho después para justificar la sustitución de los reactivos de ensayo por los reactivos objetivos de opción múltiple.

A la vuelta del siglo empezaron a aparecer las primeras pruebas estandarizadas para medir los resultados de la instrucción escolar. Encabezadas por el trabajo de E. L. Thorndike, las pruebas empleaban principios de medición tomados del laboratorio psicológico. Entre los ejemplos se incluyen escalas para calificar la calidad de la escritura y la redacción, así como pruebas de ortografía, aritmética, cálculo y razonamiento aritmético. Después llegaron las baterías de aprovechamiento, iniciadas con la publicación de la primera edición de la Prueba de Aprovechamiento de Stanford (*Stanford Achievement Test*) en 1923. Sus autores fueron tres líderes en la elaboración de pruebas: Truman L. Kelley, Giles M. Ruch y Lewis M. Terman. Presagiando muchas características de la psicometría moderna, esta batería proporcionó medidas comparables de ejecución en diferentes materias escolares, evaluadas en términos de un solo grupo normativo.

Al mismo tiempo se acumulaban evidencias de la falta de acuerdo entre los maestros al calificar las pruebas de ensayo. Para 1930 era ampliamente reconocido que las pruebas de ensayo no sólo requerían más tiempo para los examinadores y examinados, sino que también arrojaban resultados menos confiables que el "nuevo tipo" de reactivos objetivos.[4] En la medida en que estos últimos llevaron al uso creciente de las pruebas estandarizadas de aprovechamiento, hubo un énfasis mayor en el diseño

[4] La investigación relacionada con la relativa eficacia de los reactivos de ensayo y objetivos se trata en el capítulo 17, en la parte que aborda la utilización educativa de las pruebas

de reactivos para probar la comprensión y la aplicación del conocimiento y otros objetivos educativos más amplios. La década de los treinta también presenció la introducción de las máquinas para calificar exámenes a las cuales podían adaptarse fácilmente las nuevas pruebas objetivas.

Otro desarrollo digno de mención fue el establecimiento en los Estados Unidos de programas estatales, regionales y nacionales de evaluación. Probablemente el más conocido sea el de la Junta Universitaria de Exámenes de Admisión a la Universidad (*College Entrance Examination Board*, CEEB). Establecido a principios de siglo para reducir la duplicación de los exámenes de admisión presentados por los estudiantes de reciente ingreso a la universidad, el programa ha experimentado cambios profundos en sus procedimientos de evaluación y en el número y la naturaleza de las universidades participantes, cambios que reflejan los adelantos tanto en la aplicación de pruebas como en la educación. En 1947, las funciones de evaluación del CEEB se combinaron con las de la Corporación Carnegie y las del Consejo Estadounidense de Educación para formar el Servicio de Pruebas Educativas (*Educational Testing Service*, ETS). En los años posteriores, el ETS asumió la responsabilidad de un número creciente de programas de evaluación al servicio de universidades, escuelas profesionales, dependencias gubernamentales y otras instituciones. Debe hacerse mención del Programa de Evaluación de Universidades Estadounidenses (*American College Testing Program*), establecido en 1959 para seleccionar a los aspirantes a las universidades no incluidas en el programa CEEB y de varios programas nacionales de evaluación para premiar a los estudiantes talentosos.

Las pruebas de aprovechamiento no sólo se utilizan con propósitos educativos, sino también en la selección de solicitantes de empleos en la industria y el gobierno. Ya citamos el uso sistemático en el servicio civil de exámenes en el imperio chino desde aproximadamente el año 150 a.C. (Bowman, 1989). En los tiempos modernos, la selección de empleados gubernamentales por medio de exámenes fue introducida en los países europeos a finales del siglo XVIII y comienzos del XIX. En 1883, la Comisión del Servicio Civil de los Estados Unidos (*U.S. Civil Service Commission*) estableció el uso de exámenes competitivos como procedimiento regular (Kavruck, 1956). Las técnicas de elaboración de pruebas creadas durante y antes de la Primera Guerra Mundial fueron introducidas en el programa de evaluación del Servicio Civil de los Estados Unidos con el nombramiento de L. J. O'Rourke como director de la recién establecida división de investigación en 1922. En la actualidad, este trabajo lo realiza un sofisticado equipo de investigación de la Oficina de Administración de Personal de los Estados Unidos (*U.S. Office of Personnel Management*).

En la medida en que crecía la participación de psicólogos con preparación en psicometría en la formulación de pruebas estandarizadas de aprovechamiento aumentaba la semejanza de sus aspectos técnicos con el de los tests de inteligencia y de aptitud. Los procedimientos para la elaboración y evaluación de todas esas nuevas pruebas tenían mucho en común. Los crecientes esfuerzos por preparar pruebas de aprovechamiento que midieran la consecución de amplias metas educativas, en contraste con el recuerdo de hechos triviales, también hizo que el contenido de las pruebas de aprovechamiento se asemejara al de los tests de inteligencia. En la actualidad, la diferencia entre los dos instrumentos corresponde principalmente al grado de especificidad del contenido y el grado en que el instrumento presupone una instrucción previa.

EVALUACIÓN DE LA PERSONALIDAD

Otra área de interés de las pruebas psicológicas son los aspectos afectivos o no intelectuales de la conducta, los que revisaremos del capítulo 13 al 16. Los instrumentos diseñados con este propósito suelen conocerse como tests de personalidad, aunque muchos psicólogos prefieren emplear el término "personalidad" en un sentido más amplio para referirse al individuo en su totalidad. De acuerdo con esto, tanto los rasgos intelectuales como los no intelectuales deberían agruparse bajo dicho rubro; sin embargo, en la terminología psicométrica es más común el uso de la expresión "test de personalidad" para referirse a la medición de características como los estados emocionales, las relaciones interpersonales, la motivación, los intereses y las actitudes.

Kraepelin

El uso que hizo Kraepelin de la prueba de asociación libre con pacientes psiquiátricos es un antecedente de los tests de personalidad. En esta prueba, se presentan al examinado palabras estímulo especialmente seleccionadas y se le pide que responda a cada una con la primera palabra que le venga a la mente. Kraepelin (1892) también utilizó esta técnica para estudiar los efectos psicológicos de la fatiga, el hambre y las drogas, y concluyó que todos esos agentes incrementan la frecuencia relativa de asociaciones superficiales. Sommer (1894), que también trabajó en la última década del siglo XIX, sugirió que la prueba de asociación libre podría utilizarse para diferenciar formas de trastorno mental. Después, la técnica ha sido utilizada con distintos propósitos de evaluación y aún sigue empleándose. Debe mencionarse el trabajo de Galton, Pearson y Cattell en la preparación de cuestionarios estandarizados y escalas de calificación. Aunque originalmente estaban destinados a otros propósitos, estos procedimientos fueron utilizados para elaborar algunos de los tests de personalidad que hoy son más comunes.

Sommer

Woodworth

El prototipo de cuestionario de personalidad, o *inventario autodescriptivo* (capítulo 13), es la Hoja de Datos Personales (*Personal Data Sheet*) creada por Woodworth durante la Primera Guerra Mundial (DuBois, 1970; Franz, 1919, pp. 171–176; L. R. Goldberg, 1971; Symonds, 1931, capítulo 5). El cuestionario fue diseñado como una herramienta de selección para identificar a los individuos gravemente perturbados que deberían ser excluidos del servicio militar. El cuestionario constaba de una serie de preguntas que versaban sobre síntomas psicopatológicos comunes y en las que los individuos respondían sobre sí mismos. Se obtenía una puntuación total, contando el número de síntomas indicados. Este instrumento no se terminó y no pudo emplearse a tiempo, antes de que terminara la guerra, pero inmediatamente después se prepararon formas para uso civil, incluyendo una forma especial para aplicar a niños. Más aún, la Hoja de Datos Personales de Woodworth sirvió como modelo para la mayor parte de los inventarios de ajuste emocional. En algunos de estos cuestionarios se hacía un intento por subdividir el ajuste emocional en formas más específicas, como el ajuste al hogar, el escolar y el vocacional. Otros instrumentos se concentraban en un área más estrecha de conducta o en respuestas más claramente sociales, como las de dominancia-sumisión en el trato personal. El último avance fue la elaboración de instrumentos para cuantificar la expresión de actitudes e intereses (capítulo 14), que también se basaban principalmente en las técnicas de cuestionario.

Otro método de medición de la personalidad se encuentra en la aplicación de *tests situacionales* y de *ejecución* (capítulo 16), en las que el examinado debe realizar una

tarea cuyo propósito a menudo está encubierto. La mayor parte simula con mucho realismo situaciones de la vida cotidiana. La primera aplicación extensa de las técnicas está en las pruebas elaboradas por Hartshorne, May y colaboradores a finales de los veinte y principios de los treinta (1928, 1929, 1930). Esta serie, estandarizada para escolares, se interesaba en conductas tales como copiar, mentir, robar, cooperar y persistir. Era posible obtener puntuaciones cuantitativas objetivas en numerosos instrumentos específicos. Otro ejemplo, éste para los adultos, se encuentra en la serie de tests situacionales preparados durante la Segunda Guerra Mundial por el Programa de Evaluación de la Oficina de Servicios Estratégicos (*Office of Strategic Services*, OSS, 1948). Estos tests se interesaban en la conducta emocional y social sutil y relativamente compleja y su aplicación requería de condiciones más bien elaboradas y de personal capacitado, además de que la interpretación de las respuestas era relativamente subjetiva.

Las *técnicas proyectivas* (capítulo 15) representan el tercer método de estudio de la personalidad que ha mostrado un crecimiento notable, en especial entre los clínicos. Estos instrumentos presentan al cliente un estímulo no muy estructurado, lo que permite una considerable libertad en su solución. La suposición que fundamenta este método es que el individuo proyectará mediante el estímulo su estilo característico de respuesta. Como en los tests situacionales o de ejecución, el propósito de las técnicas proyectivas está más o menos encubierto, lo que reduce la posibilidad de que el individuo cree deliberadamente una impresión deseable. La prueba de asociación libre, que ya citamos, es una de las primeras técnicas proyectivas. Los tests de frases incompletas también se han utilizado de esta manera. Otras tareas que aparecen en las técnicas proyectivas incluyen el dibujo, el arreglo de juguetes para crear una escena, la dramatización extemporánea y la interpretación de manchas de tinta.

Todos los tests de personalidad disponibles presentan ciertas dificultades prácticas y teóricas. Cada método tiene sus propias ventajas y desventajas. En conjunto, los tests de personalidad han quedado detrás de los de habilidad en cuanto a logros prácticos, pero esa falta de progreso no puede atribuirse a un esfuerzo insuficiente. La investigación sobre la medición de la personalidad ha alcanzado proporciones impresionantes desde 1950, y muchos instrumentos ingeniosos y mejoras técnicas están en investigación. Lo que explica el lento avance en el área son más bien las dificultades específicas que se encuentran en la medición de la personalidad.

A partir de la investigación actual con los tests de personalidad están surgiendo dos tendencias unificadoras importantes (véase Anastasi, 1985b, 1992a, 1993; Digman, 1990; L. R. Goldberg, 1993; Simon, 1994). Primera, cada vez hay más evidencias de la influencia recíproca de los rasgos afectivos (de "personalidad") y cognoscitivos (de "habilidad") tanto en el desempeño de tareas como en el desarrollo conductual. Se ha llegado a la conclusión de que la distinción tradicional entre los dos rasgos ha sido impuesta artificialmente por razones de conveniencia en la descripción y medición de diferentes aspectos de la conducta. Segunda, el análisis teórico de la naturaleza y composición de la personalidad apoya la integración de los rasgos cognoscitivos y afectivos en un modelo amplio de la actividad humana que incluye todas las formas de conducta. Este modelo relaciona la investigación básica de los rasgos intelectuales (capítulo 11) y los afectivos (capítulo 13).

SEGUNDA PARTE

Principios técnicos y metodológicos

Normas y significado de las puntuaciones de los tests

En la segunda parte, compuesta por los capítulos 3 a 7, encontrará una introducción a los conceptos básicos y la metodología que se requieren para entender las pruebas psicológicas y para la adecuada interpretación de sus resultados. Los siguientes capítulos abordan temas como las normas, la confiabilidad, la validez, el análisis de reactivos y el diseño de los tests. Este capítulo trata sobre el desarrollo y uso de las normas y otros procedimientos que facilitan la interpretación de las puntuaciones. En ausencia de datos interpretativos adicionales, la sola puntuación cruda de cualquier prueba psicológica carece de significado. Decir que un individuo ha resuelto correctamente 15 problemas de una prueba de razonamiento aritmético, identificado 34 palabras de una prueba de vocabulario o armado un objeto mecánico en 57 segundos, transmite poca o ninguna información sobre su posición en cualquiera de estas funciones. Tampoco los porcentajes brindan una solución satisfactoria al problema de interpretar las puntuaciones de los instrumentos; por ejemplo, una puntuación de 65 por ciento de aciertos en una prueba de vocabulario puede ser equivalente al 30 en otra y al 80 por ciento en una más. Obviamente, el grado de dificultad de los reactivos que componen cada prueba determinará el significado del resultado. Al igual que las puntuaciones crudas, los porcentajes sólo pueden interpretarse en un marco de referencia claramente definido y uniforme.

En los instrumentos psicológicos, las puntuaciones suelen interpretarse haciendo referencia a *normas* que representan el desempeño de la muestra de estandarización en la prueba; es decir, las normas se establecen empíricamente al determinar lo que hacen en la prueba los miembros de un grupo representativo. La puntuación transformada de cualquier individuo se refiere entonces a la distribución de las puntuaciones obtenidas por la muestra de estandarización para descubrir qué lugar ocupa en esa distribución. ¿Coincide el resultado con el desempeño promedio del grupo de estanda-

rización? ¿Está ligeramente por debajo del promedio o se encuentra cerca del extremo superior de la distribución?

Para averiguar la posición exacta del individuo en relación con la muestra de estandarización, la puntuación cruda se convierte en una medida relativa. Estas puntuaciones transformadas cumplen un doble propósito; primero, indican la posición del individuo en relación con la muestra normativa, lo que permite la evaluación de su desempeño en comparación con otras personas. Segundo, proporcionan medidas equivalentes que permiten la comparación directa del desempeño del individuo en pruebas diferentes; por ejemplo, si una chica obtiene una puntuación cruda de 40 en una prueba de vocabulario y una de 22 en una prueba de razonamiento aritmético, esto nada nos dice acerca de su desempeño relativo en las dos pruebas: ¿es mejor en vocabulario o en aritmética, o bien igualmente buena en ambas materias? Como las puntuaciones crudas de diferentes pruebas suelen expresarse en unidades distintas resulta imposible su comparación directa. El nivel de dificultad de cada prueba también afecta esta comparación. Por otro lado, las puntuaciones derivadas pueden expresarse en las mismas unidades y referirlas a muestras normativas iguales o muy similares en distintas pruebas, lo que permite comparar la ejecución relativa en funciones muy diferentes.

Aunque hay varias formas de convertir las puntuaciones crudas para satisfacer los dos objetivos planteados, las puntuaciones derivadas se expresan esencialmente en dos formas principales: (1) el nivel de desarrollo alcanzado, o (2) la posición relativa dentro de un grupo específico. Estudiaremos ambas puntuaciones y algunas de sus variantes en otras secciones del capítulo, pero antes es necesario examinar algunos conceptos elementales de estadística que fundamentan el desarrollo y la utilización de las normas. La siguiente sección aclara el significado de ciertas medidas estadísticas comunes, y proporciona ejemplos simplificados de los cálculos sin pretender con ello brindar una capacitación en los métodos estadísticos. Para los detalles de cálculo y los procedimientos a seguir en la aplicación práctica de esas técnicas, acuda el lector a cualquier texto actualizado de estadística (por ejemplo, D. C. Howell, 1997; Runyon y Haber, 1991; West, 1991). En la actualidad, es común que se reconozca la necesidad de que no sólo los usuarios de los tests cuenten con conocimientos básicos de la metodología de la estadística, sino también cualquiera que desee comprender los informes de investigación publicados en cualquier área de la psicología (L. S. Aiken, West, Sechrest y Reno, 1990; Anastasi, 1991; Lambert, 1991; S. T. Meier, 1993).

CONCEPTOS ESTADÍSTICOS

Un propósito fundamental de la estadística es organizar y resumir los datos cuantitativos para facilitar su comprensión. Una lista de 1 000 puntuaciones no sólo resulta abrumadora, sino que transmite poco significado. El primer paso para poner orden en semejante caos de datos es tabular las puntuaciones en una *distribución de frecuencias*, como se ilustra en la tabla 3.1, que agrupa las puntuaciones en intervalos de clase, y marca cada una en el intervalo adecuado. Cuando se han anotado todas las puntuaciones se cuentan las marcas para encontrar la frecuencia o el número de casos en cada intervalo de clase. La suma de esas frecuencias será igual a N, el número total de casos en el grupo. La tabla 3.1 muestra las puntuaciones obtenidas por 1 000 universi-

Tabla 3.1

Distribución de frecuencias de las puntuaciones de 1 000 estudiantes universitarios en una prueba de aprendizaje de códigos

Intervalo de clase	Frecuencia
52–55	1
48–51	1
44–47	20
40–43	73
36–39	156
32–35	328
28–31	244
24–27	136
20–23	28
16–19	8
12–15	3
8–11	2
	N = 1 000

(Datos tomados de Anastasi, 1934, pág. 34)

tarios en una prueba de aprendizaje de códigos en la que un conjunto de palabras artificiales o sílabas sin sentido debía ser sustituido por otro. Las puntuaciones crudas indicaban el número de sílabas correctas sustituidas en un ensayo de dos minutos, con un rango de ocho a 52. Se agruparon en intervalos de clase de cuatro puntos, que iban de 52–55 en la parte superior de la distribución a 8–11 en el extremo inferior. La columna de frecuencia revela que dos estudiantes obtuvieron puntuaciones entre ocho y 11, tres entre 12 y 15, ocho entre 16 y 19, etcétera.

La información que proporciona una distribución de frecuencias también puede representarse en forma gráfica en una curva de distribución. La figura 3.1 muestra de esa manera los datos de la tabla 3.1. En la línea base, el eje horizontal, se encuentran las puntuaciones agrupadas en intervalos de clase; en el eje vertical se localizan las frecuencias o números de casos que caen en cada intervalo. La gráfica se presenta de dos formas, ambas de uso común. En el *histograma*, la altura de la columna levantada sobre cada intervalo corresponde al número de sujetos localizados ahí. Podemos imaginar que cada individuo se para en los hombros de otro para formar la columna. En el *polígono de frecuencia*, un punto en el centro del intervalo de clase y conectado a la frecuencia apropiada indica el número de sujetos en cada intervalo. Los puntos sucesivos se unen luego con líneas rectas.

A excepción de algunas irregularidades menores, la distribución que se muestra en la figura 3.1 recuerda la *curva normal*. En la figura 3.2 se presenta una curva normal perfecta determinada matemáticamente. Esta curva tiene importantes propiedades matemáticas y brinda la base para muchos análisis estadísticos; sin embargo, para nuestros propósitos actuales sólo mencionaremos algunos de esos rasgos. En esencia, la curva indica que el mayor número de casos se agrupa en el centro de la distribución, y que al acercarse a los extremos va disminuyendo gradualmente en ambas direcciones.

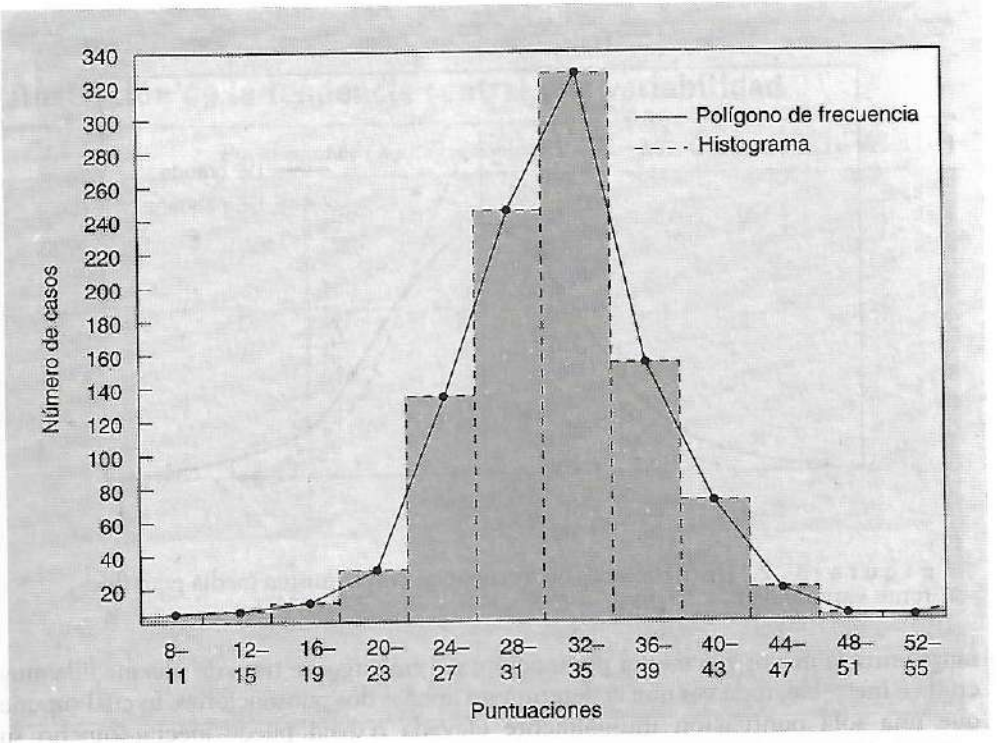


Figura 3.1. Curvas de distribución: polígono de frecuencia e histograma.
(Datos de la tabla 3.1)

La curva es simétrica y tiene un único pico en el centro. Casi todas las distribuciones de rasgos humanos, estatura, peso, aptitudes y características de personalidad, se aproximan a la curva normal, y, en general, entre más grande sea el grupo más parecida resultará la distribución a la curva teórica normal.

Un grupo de puntuaciones también puede describirse en términos de alguna medida de *tendencia central*, que ofrece la puntuación más común o representativa del desempeño de todo el grupo. La más conocida de esas medidas es el promedio, que técnicamente se llama *media* (M) y se obtiene al sumar todas las puntuaciones y dividir la suma entre el número de casos (*N*). Otra medida es la *moda*, o puntuación más frecuente. En una distribución de frecuencias, la moda es el punto central del intervalo de clase con la mayor frecuencia. De acuerdo con lo anterior, en la tabla 3.1 la moda cae a la mitad del intervalo 32–35 y es igual a 33.5. Observe que esta puntuación corresponde al punto más alto en la curva de distribución de la figura 3.1. Otra medida de tendencia central es la *mediana*, el punto más cercano al centro de una distribución de puntuaciones ordenadas por tamaño. La mediana es el punto que divide la distribución por encima y por debajo del cual cae la mitad de los casos.

Las medidas de *variabilidad*, de la extensión de las diferencias individuales alrededor de la tendencia central, proporcionan otra descripción del conjunto de puntuaciones de un test. La forma más evidente y familiar de mostrar la variabilidad es el

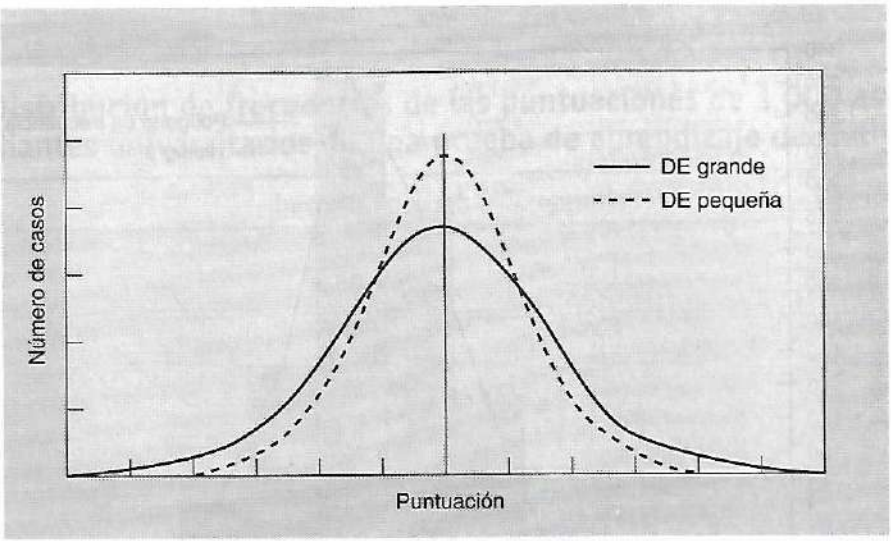


Figura 3.2. Distribuciones de frecuencias con la misma media pero diferente variabilidad.

rango entre la mayor y la menor puntuación; sin embargo, se trata de una medida muy cruda e inestable, toda vez que la determinan apenas dos puntuaciones, lo cual supone que una sola puntuación inusualmente elevada o baja puede afectar mucho su tamaño. Un método más preciso para medir la variabilidad se basa en la diferencia entre la puntuación de cada individuo y la media del grupo.

En este punto resulta útil estudiar el ejemplo de la tabla 3.2, en el que se calcularon las diversas medidas consideradas para 10 casos. Aunque en la práctica real suelen manejarse grupos más grandes, en este caso la elección de un grupo tan pequeño obedece a la necesidad de simplificar la explicación. La tabla 3.2 también permite introducir algunos símbolos estadísticos que debe aprender, ya que los emplearemos más adelante. De manera convencional, designamos las puntuaciones crudas originales con una X mayúscula, la *x* minúscula se emplea para referirse a las desviaciones de cada puntuación respecto a la media del grupo y la letra griega Σ significa "la suma de". Se observa que en la primera columna de la tabla 3.2 aparecen los datos para el cálculo de la media (que en este caso es 40) y la mediana (que es igual a 40.5 y se encuentra entre 40 y 41) por encima y por debajo de la cual se encuentran respectivamente cinco casos (50 por ciento). En un grupo tan pequeño no tiene sentido buscar la moda, porque los casos no muestran un agrupamiento inequívoco en ninguna puntuación; de cualquier modo, la puntuación de 41 representaría técnicamente a la moda, porque la obtuvieron dos individuos, en tanto que el resto de los resultados ocurrió una sola vez.

La segunda columna muestra qué tanto se desvía cada puntuación por encima o por debajo de la media de 40. La suma de las desviaciones siempre será igual a cero, porque las desviaciones positivas y las negativas alrededor de la media necesariamente se equilibran o cancelan (+20 –20 = 0). Si ignoramos los signos podemos promediar las desviaciones absolutas y obtener una medida de la desviación promedio de cada sujeto de la media del grupo. Aunque de cierto valor descriptivo, en la práctica no se

Tabla 3.2

Ilustración de la tendencia central y la variabilidad

	Puntuaciones (X)	Desviación ($x = X - M$)	Cuadrado de las desv. (x^2)
50% de los casos	48	+8	64
	47	+7	49
	43	+3	9
	41	+1	1
Mediana = 40.5 →	41	+1	1
		+20	
	40	0	0
50% de los casos	38	−2	4
	36	−4	16
	34	−6	36
	32	−8	64
		−20	
	$\Sigma X = 400$		$\Sigma x^2 = 244$

$$M = \frac{\Sigma X}{N} = \frac{400}{10} = 40$$

$$\text{Varianza} = \sigma^2 = \frac{\Sigma x^2}{N} = \frac{244}{10} = 24.40$$

$$DE \text{ o } \sigma = \sqrt{\frac{\Sigma x^2}{N}} = \sqrt{24.40} = 4.9$$

Nota. Los símbolos Σ y σ son la mayúscula y la minúscula de la misma letra griega, pronunciada "sigma". En muchos textos de estadística, la *DE* se refiere a la desviación estándar de la muestra de la que se obtuvieron los datos, mientras que σ se refiere al estimado de la desviación estándar en la población de la que se extrajo la muestra.

emplea la "desviación promedio" porque la eliminación arbitraria de los signos hace inadecuado su uso en posteriores análisis matemáticos.

La *desviación estándar* (simbolizada por *DE* o σ) es una medida de variabilidad mucho más útil en la que los signos negativos se eliminan al elevar al cuadrado cada desviación, procedimiento seguido en la última columna de la tabla 3.2. Se conoce como *varianza* o *media de los cuadrados de las desviaciones* a la suma de esta columna dividida entre el número de casos[1] $\left(\frac{\Sigma x^2}{N}\right)$. La varianza ha demostrado ser de gran utilidad para separar la contribución de diferentes factores a las diferencias individuales en el desempeño de una prueba; sin embargo, para los propósitos presentes lo que más nos interesa es la *DE* que, como se muestra en la tabla 3.2, es la raíz cuadrada de la varianza. Esta medida suele emplearse al comparar la variabilidad de distintos grupos; por ejemplo, la figura 3.2 presenta dos distribuciones que coinciden en la media pero difieren en la variabilidad. La distribución con la mayor variabili-

[1] Los cálculos ilustrados en este capítulo tienen que ver con la *estadística descriptiva*, que se refiere a la muestra realmente medida; en la *estadística inferencial*, la *N* es reemplazada por *N*-1 para estimar los valores correspondientes a la población a partir de los datos de la muestra. Entre menor sea la muestra, mayor será la diferencia entre los valores de la muestra y los de la población. Para más explicaciones, véase cualquier libro actualizado de estadísticas (por ejemplo, Comrey y Lee, 1992).

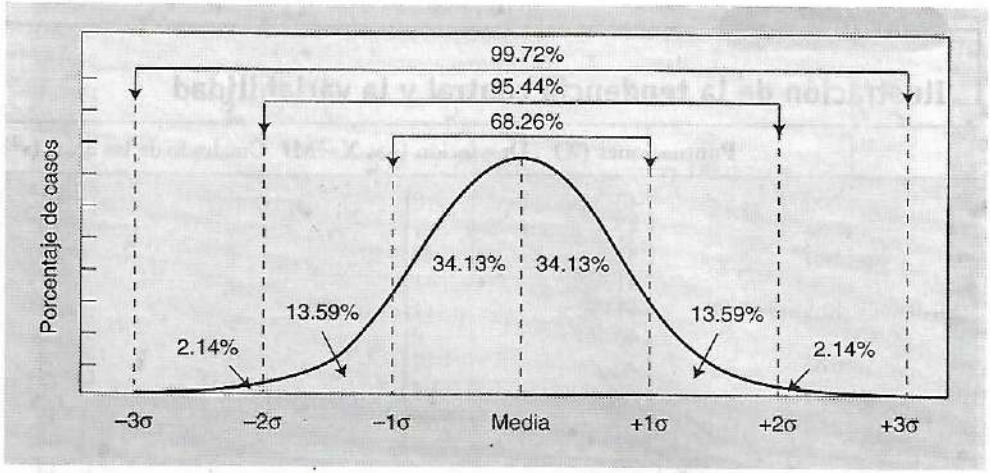


Figura 3.3. Distribución de los porcentajes de casos de la curva normal.

dad individual produce una *DE* más grande que la que tiene menores diferencias individuales. Al evaluar la ejecución relativa de dos grupos, deben compararse tanto las *DE* como las medias. Si se encuentra que la variabilidad de los grupos difiere, puede suponerse que son disímiles en la proporción de puntuaciones altas, bajas o en ambas, independientemente de las diferencias entre las medias. En la actualidad se dispone de procedimientos que permiten combinar los efectos de las diferencias entre las medias y las diferencias en la *DE* (véase, por ejemplo, Feingold, 1995).

La *DE* también proporciona la base para expresar en términos de normas las puntuaciones del individuo en diferentes pruebas, como se mostrará en la sección de calificaciones estándares. La interpretación de la *DE* resulta especialmente clara al aplicarse a una curva de distribución normal o aproximadamente normal, en la que hay una relación exacta entre la *DE* y la proporción de casos, como se ve en la figura 3.3. En la línea horizontal de la curva se han marcado las distancias que representan una, dos y tres desviaciones normales por encima y por debajo de la media; por ejemplo, en el caso presentado en la tabla 3.2, la media correspondería a una puntuación de 40, $+1\sigma$ correspondería a 44.9 (es decir, 40 + 4.9); $+2\sigma$ a 49.8 (es decir, 40 + 2 × 4.9), etc. En la curva normal el porcentaje de casos que cae entre la media y $+1\sigma$ es 34.13 por ciento. Como la curva es simétrica, también se encuentra 34.13 por ciento de los casos entre la media y -1σ, de modo que entre $+1\sigma$ y -1σ a ambos lados de la media se encuentra el 68.26 por ciento de los casos. Casi todos los casos (99.72 por ciento) caen dentro de $\pm 3\sigma$ de la media. Estas relaciones son particularmente importantes en la interpretación de las puntuaciones estándares y los percentiles que veremos adelante.

NORMAS DE DESARROLLO

Las puntuaciones de las pruebas adquieren significado si indican qué tanto ha progresado el individuo en el patrón de desarrollo normal. Así, puede decirse que un niño de ocho años tiene una edad mental (EM) de 10 si su ejecución en una prueba de in-

teligencia es tan buena como la de un niño de 10; la EM de 10 también se asignaría a un adulto retardado que mostrara el mismo nivel de ejecución. En un contexto diferente, puede decirse que un chico de cuarto grado ha alcanzado la norma de sexto en una prueba de lectura y la de tercero en una de aritmética. Otros sistemas de desarrollo utilizan descripciones cualitativas de la conducta en funciones específicas, como las actividades sensoriomotoras o la formación de conceptos. Como quiera que se expresen, las puntuaciones basadas en normas de desarrollo tienden a ser burdas en términos de la psicometría, y no se prestan a un tratamiento estadístico preciso; no obstante, resultan muy atractivas para propósitos descriptivos, en especial en el estudio clínico intensivo de individuos y para ciertos propósitos de investigación.

Edad mental. En el capítulo 2 dijimos que el término "edad mental" se popularizó gracias a las traducciones y adaptaciones de las escalas de Binet-Simon, aunque el propio Binet prefería usar el más neutral de "nivel mental". En escalas de edad como las de Binet y sus revisiones (previas a 1986), los reactivos se agrupaban en niveles cronológicos. Por ejemplo, los reactivos pasados por la mayoría de los niños de siete años de una muestra de estandarización se colocaban en el nivel de siete años, los pasados por casi todos los de ocho años se asignaban al nivel de ocho años, etc. La puntuación de un niño en la prueba correspondería entonces al mayor nivel que hubiera completado con éxito. En la práctica real, la ejecución de los individuos mostraba cierto grado de *dispersión*, lo que significa que en algunas pruebas estaban por debajo de su edad mental y pasaban otras por encima. Por esta razón se acostumbraba calcular la *edad basal*, es decir, la mayor edad en la que todas las pruebas se pasaban. A esta edad basal se agregaban luego créditos parciales en meses por las pruebas pasadas en niveles superiores. La edad mental del niño en la prueba era la suma de la edad basal y los meses adquiridos como créditos en los niveles superiores.

Las normas de edad mental también han sido empleadas con pruebas que no se dividen en niveles de años. En ese caso, primero se determina la puntuación del niño, que puede ser el número total de aciertos obtenido en toda la prueba o bien puede basarse en el tiempo, número de errores o en alguna combinación de esas medidas. El promedio de las puntuaciones crudas de los niños de la muestra de estandarización en cada grupo de edad constituye la norma de edad para esa prueba; por ejemplo, la puntuación cruda promedio de los niños de ocho años representa la norma para esa edad. Si la calificación de un individuo es igual a la puntuación cruda promedio de los ocho años, entonces su edad mental en la prueba es de ocho años. Todas las puntuaciones crudas de dicha prueba pueden transformarse de manera similar haciendo referencia a las normas de edad.

Advierta que la unidad de edad mental no permanece constante con la edad, sino que tiende a disminuir con los años; por ejemplo, un niño que a la edad de cuatro tiene un año de retardo a los 12 tendrá aproximadamente tres años de retardo. Un año de desarrollo mental de los tres a los cuatro años equivale a tres años de desarrollo de los nueve a los 12. Como el desarrollo intelectual avanza más rápidamente a edades tempranas y disminuye a medida que el individuo se aproxima a su límite de maduración, la unidad de edad mental se reduce también con la edad. Esta relación se aprecia con más facilidad si se piensa en la estatura del individuo expresada en términos de "edad de estatura". La diferencia, en centímetros, entre una edad de estatura de tres y cuatro años será mayor que entre una de 10 y 11. Debido a la disminución progresiva

de la unidad de EM, un año de aceleración o retardo, digamos, la edad de cinco representa una desviación mayor de la norma que un año de aceleración o retardo a los 10.

Grados equivalentes. En las pruebas de aprovechamiento académico, los resultados suelen interpretarse como equivalentes a grados, práctica comprensible porque las pruebas se emplean en medios escolares. Describir el aprovechamiento de un alumno como equivalente al desempeño de séptimo grado en ortografía, de octavo en lectura y de quinto en aritmética tiene el mismo atractivo popular que el uso de la edad mental en los tests tradicionales de inteligencia.

Las normas de grado se encuentran calculando la puntuación cruda promedio de los niños de cada grado. De esta manera, si en la muestra de estandarización los chicos de cuarto resuelven correctamente un promedio de 23 problemas en una prueba de aritmética, entonces la puntuación cruda de 23 corresponde a un grado equivalente de cuatro. Los equivalentes de grados intermedios, que representan fracciones de grado, en general se encuentran por interpolación, aunque también pueden obtenerse examinando directamente a los niños en momentos diferentes del año escolar. Como éste cubre 10 meses, pueden expresarse como decimales; por ejemplo, 4.0 se refiere al desempeño promedio al inicio del cuarto grado (aplicando la prueba al empezar el curso), 4.5 al desempeño promedio a la mitad del grado (aplicando la prueba a mediados del año escolar), etcétera.

A pesar de su popularidad, las normas de grado tienen varios defectos. Primero, el contenido de la instrucción varía de un grado a otro, de ahí que las normas de grado sólo sean apropiadas para las materias enseñadas a lo largo de todos los niveles escolares cubiertos por la prueba. Por lo general no son aplicables en el nivel de secundaria, pues muchas materias se cursan sólo uno o dos años; sin embargo, incluso en las materias enseñadas en cada grado el énfasis puesto en diferentes temas puede variar, por lo que el progreso puede ser más rápido en una materia que en otra. En otras palabras, las unidades de grado son desiguales, y tales desigualdades ocurren de manera irregular en diferentes áreas temáticas.

Las normas de grado también pueden ser objeto de malas interpretaciones a menos que el usuario tenga siempre presente la manera en la que se derivaron; por ejemplo, si una niña de cuarto obtiene un grado equivalente de 6.9 en aritmética, eso *no* significa que domine los procesos aritméticos enseñados en sexto. Aunque es indudable que su calificación obedece a una ejecución superior en la aritmética de cuarto grado, no debe suponerse que cumple los requisitos para la aritmética de séptimo. Por último, las normas de grado tienden a ser consideradas en forma errónea como criterios de desempeño; por ejemplo, una maestra de sexto puede suponer que todos sus alumnos deben encontrarse alrededor de las normas del sexto grado en las pruebas de aprovechamiento. Esta concepción errónea no es sorprendente cuando se utilizan normas de grado. Con todo, las diferencias individuales observadas en cualquier grado son tan grandes, que el rango de las puntuaciones de una prueba de aprovechamiento inevitablemente se extiende a lo largo de varios grados.

Escalas ordinales. Otra aproximación a las normas de desarrollo proviene de la investigación en el campo de la psicología infantil. Las observaciones empíricas del desarrollo conductual de infantes y niños pequeños lleva a la descripción de la conducta característica de esas edades sucesivas en funciones como la locomoción, la discriminación sensoriomotora, la comunicación lingüística y la formación de concep-

tos. Un ejemplo precoz se encuentra en el trabajo de Gesell y sus colegas en Yale (Ames, 1937; Gesell y Amatruda, 1947; Halverson, 1933; Knobloch y Pasamanick, 1974). Los Programas de Desarrollo de Gesell presentan, en meses, el nivel aproximado de desarrollo que ha alcanzado el niño en cuatro áreas principales de conducta, a saber, motora, adaptativa, lenguaje y personal-social. Estos niveles se encuentran comparando la conducta del niño con la que es distintiva de ocho edades claves que van de las cuatro semanas a los 36 meses.

Gesell y sus colaboradores subrayaban el patrón secuenciado del desarrollo conductual temprano. Citaban evidencias considerables de las uniformidades de las secuencias del desarrollo y la progresión ordenada de los cambios en la conducta; por ejemplo, en las reacciones del niño ante un objeto pequeño colocado frente a él se encuentra una secuencia cronológica característica en la fijación visual y los movimientos de mano y dedos. El uso de la mano en intentos burdos de prensión palmar ocurren a una edad anterior al uso del pulgar en oposición a la palma; esta prensión es a su vez seguida por el uso más eficiente de los dedos pulgar e índice a manera de tenazas al asir el objeto. Este patrón secuenciado se observa también al caminar, subir escaleras y en la mayor parte del desarrollo sensoriomotor de los primeros años. Las escalas desarrolladas dentro de este marco son ordinales en el sentido de que las etapas del desarrollo siguen un orden constante, en el que cada etapa presupone el dominio de la conducta que es característica de las etapas previas.[2]

En los sesenta se observó el resurgimiento del interés en las teorías del desarrollo del psicólogo infantil suizo Jean Piaget (véase Flavell, 1963; Ginsburg y Opper, 1969; D. R. Green, Ford y Flamer, 1971). La investigación de Piaget se centró en el desarrollo de los procesos cognoscitivos de la infancia a la adolescencia. Se interesaba en conceptos específicos más que en habilidades generales. La permanencia del objeto es un ejemplo de dichos conceptos o esquemas. En este caso, el niño es consciente de la identidad y la existencia continua de objetos cuando los ve desde ángulos diferentes o cuando salen del alcance de su vista. Otro concepto muy estudiado es el de la conservación, el reconocimiento de que un atributo permanece constante a pesar de los cambios en su apariencia, como cuando la misma cantidad de líquido se vierte en recipientes de forma distinta o cuando se colocan varillas de la misma longitud en órdenes diferentes.

Las tareas piagetianas han sido exhaustivamente empleadas por los psicólogos del desarrollo en su labor de investigación, y algunas han sido organizadas en escalas estandarizadas, como veremos en el capítulo 9 (Goldschmid y Bentler, 1968b; Pinard y Laurendeau, 1964; Užgiris y Hunt, 1975). De acuerdo con la aproximación de Piaget, esos instrumentos son escalas ordinales en las que alcanzar una etapa depende de haber completado las etapas anteriores en el desarrollo del concepto. Las tareas se diseñan para revelar los aspectos dominantes de cada etapa de desarrollo; sólo después se obtienen datos empíricos concernientes a las edades en las que suele alcanzarse

[2] Este uso del término "escala ordinal" difiere del que se le da en estadística, en la que una escala ordinal es simplemente aquella que permite el ordenamiento de los individuos sin conocimiento sobre la cantidad de diferencia entre ellos; en sentido estadístico, las escalas ordinales se contrastan con escalas de intervalo de unidad-igual. Las escalas ordinales del desarrollo infantil se diseñan sobre el modelo de la escala de Guttman, o simplex, en la que la ejecución exitosa en un nivel implica el dominio de todos los niveles inferiores (L. Guttman, 1944). Bart y Airasian (1974) describen una extensión del análisis de Guttman para incluir jerarquías no lineales, con referencia especial a las escalas piagetanas.

cada etapa. A este respecto, el procedimiento difiere del seguido en la elaboración de escalas de edad en las que los reactivos se eligen en primer lugar sobre la base de su diferenciación entre edades sucesivas. Aunque prosigue el interés en las contribuciones de los planteamientos de Piaget, los análisis críticos y las evaluaciones empíricas han resaltado tanto sus aportaciones como sus limitaciones (Sugarman, 1987).[3]

En suma, las escalas ordinales se diseñan para identificar la etapa alcanzada por el niño en el desarrollo de funciones conductuales específicas. Aunque los resultados pueden informarse en términos de niveles aproximados de edad, estas puntuaciones son secundarias a la descripción cualitativa de la conducta característica del niño. La ordinalidad de dichas escalas se refiere a la progresión uniforme del desarrollo por las etapas sucesivas. En la medida en que las escalas suelen proporcionar información acerca de lo que el niño verdaderamente es capaz de hacer (por ejemplo, trepar escaleras sin ayuda, reconocer la identidad de la cantidad de líquido cuando se vierte en recipientes de distintas formas), comparten rasgos importantes con las pruebas de dominio que analizamos en otra sección.

NORMAS INTRAGRUPO

Casi todos los tests estandarizados proporcionan alguna forma de normas intragrupo, con las que la ejecución del individuo se evalúa en términos del desempeño del grupo de estandarización más próximo, como sucede al comparar la calificación cruda de un niño con la de otros de la misma edad o del mismo grado escolar. Las puntuaciones intragrupo tienen un significado cuantitativo uniforme y definido y pueden emplearse adecuadamente en casi todos los análisis estadísticos.

Percentiles. Las calificaciones percentiles se expresan en términos del porcentaje de sujetos de la muestra de estandarización que caen bajo determinada puntuación cruda. Por ejemplo, si 28 por ciento resuelve correctamente menos de 15 problemas de una prueba de razonamiento aritmético, entonces la puntuación cruda de 15 corresponde al percentil 28 (P_{28}). El percentil indica la posición relativa del individuo en la muestra de estandarización. Los percentiles también pueden verse como rangos en un grupo de 100, excepto porque en éstos se acostumbra empezar a contar por la parte superior, y el mejor individuo del grupo recibe un rango de 1. Por su parte, con los percentiles se comienza por el extremo inferior, de modo que entre más bajo sea el percentil más pobre es la posición del individuo.

El percentil 50 (P_{50}) corresponde a la mediana, ya estudiada como medida de tendencia central. Los percentiles por encima de 50 representan una ejecución por encima de la mediana; los que están por debajo de 50 significan una ejecución inferior. Los percentiles 25 y 75 se conocen como el primer y el tercer puntos cuartiles (Q_1 y Q_3), porque separan los cuartos inferior y superior de la distribución. Como la mediana, proporcionan referencias convenientes para describir una distribución de puntuaciones y compararla con otras distribuciones.

---

[3] Véase el capítulo 9 para otra evaluación de la aproximación de Piaget.

Los percentiles no deben confundirse con los porcentajes, que son puntuaciones crudas que se expresan en términos del porcentaje de los reactivos correctos y el total; los percentiles son puntuaciones transformadas que se expresan en términos del porcentaje de individuos. Una puntuación cruda inferior a cualquiera de las calificaciones obtenidas en la muestra de estandarización tiene un rango percentil de cero (P_0); una superior a cualquier puntuación de la muestra de estandarización tiene una rango percentil de 100 (P_{100}); sin embargo, los percentiles no implican una puntuación cruda de cero ni una puntuación cruda perfecta.

Las rangos percentilares tienen varias ventajas: son sencillas de calcular y de comprender, incluso por personas sin capacitación técnica. Más aún, son de aplicación universal, pueden usarse lo mismo con adultos que con niños y son adecuados para cualquier prueba, sea que mida variables de aptitud o de personalidad.

Su principal desventaja está en la marcada disparidad de sus unidades, en especial en los extremos de la distribución. Si la distribución de las puntuaciones crudas se aproxima a la curva normal (como sucede con los resultados de la mayor parte de las pruebas), las diferencias en las puntuaciones crudas cercanas al centro de la distribución (o mediana) aparecen exageradas al transformarse a percentiles; por el contrario, las diferencias en las puntuaciones crudas próximas a los extremos quedan considerablemente acortadas. En la figura 3.4 puede verse esta distorsión. Como recordará, en una curva normal los casos se agrupan cerca del centro y su dispersión aumenta al aproximarse a los extremos; en consecuencia, cualquier porcentaje de casos cercanos al centro cubre una distancia más corta en la línea horizontal que el mismo porcentaje cerca de los extremos de la distribución. En la figura 3.4 puede verse fácilmente esta discrepancia en los intervalos entre los rangos percentiles (*RP*) si se compara la distancia entre un *RP* de 40 y uno de 50 con la distancia entre un *RP* de 10 y uno de 20. Todavía más sorprendente es la discrepancia entre esas distancias y la que hay en-

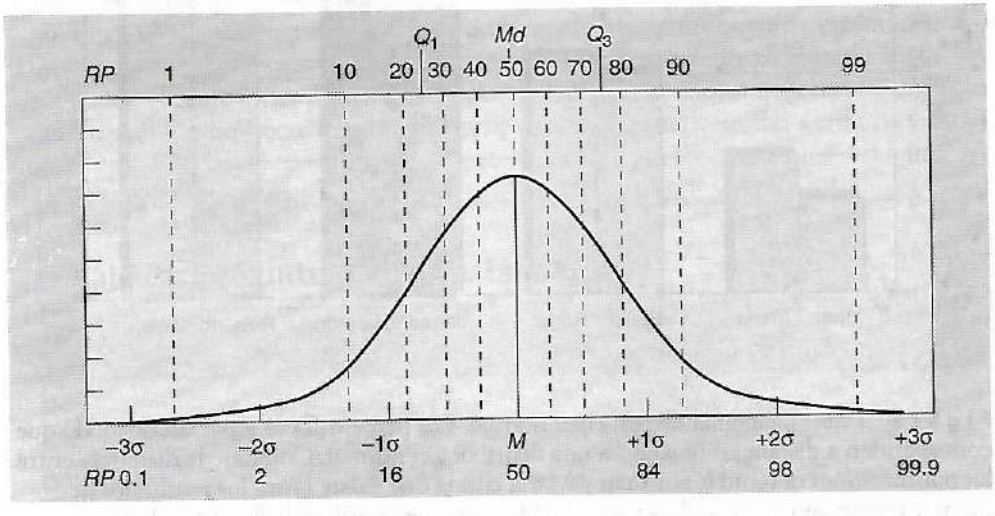


Figura 3.4. Rangos percentilares en una distribución normal (Q es igual a cuartil).

tre un *RP* de 10 y un *RP* de 1 (en una curva normal derivada matemáticamente, el percentil cero se alcanza en el infinito, por lo que no puede mostrarse en la gráfica).

La misma relación puede verse desde la dirección opuesta si se examinan los rangos percentilares correspondientes a iguales distancias σ de la media de una curva normal. Esos rangos se presentan bajo la gráfica de la figura 3.4, según la cual la diferencia percentilar entre la media y +1σ es 34 (84 – 50), y entre +1σ y +2σ es de sólo 14 (98 – 84).

Resulta evidente que los percentiles muestran la posición relativa de cada individuo en la muestra normativa pero no la magnitud de la diferencia entre puntuaciones; sin embargo, si se grafican en una carta de probabilidad aritmética, las puntuaciones percentilares también brindan una imagen correcta de las diferencias entre las puntuaciones. La carta de probabilidad aritmética es una gráfica con secciones transversales en la que las líneas verticales muestran el mismo espaciamiento que los puntos percentilares de una distribución normal (como en la figura 3.4), mientras que el espaciamiento entre las líneas horizontales es uniforme, o a la inversa (como en la figura 3.5). Las *gráficas de percentiles normales* pueden usarse para mostrar las puntuaciones de diferentes individuos en la misma prueba o los resultados de la misma persona en diferentes pruebas. En cual-

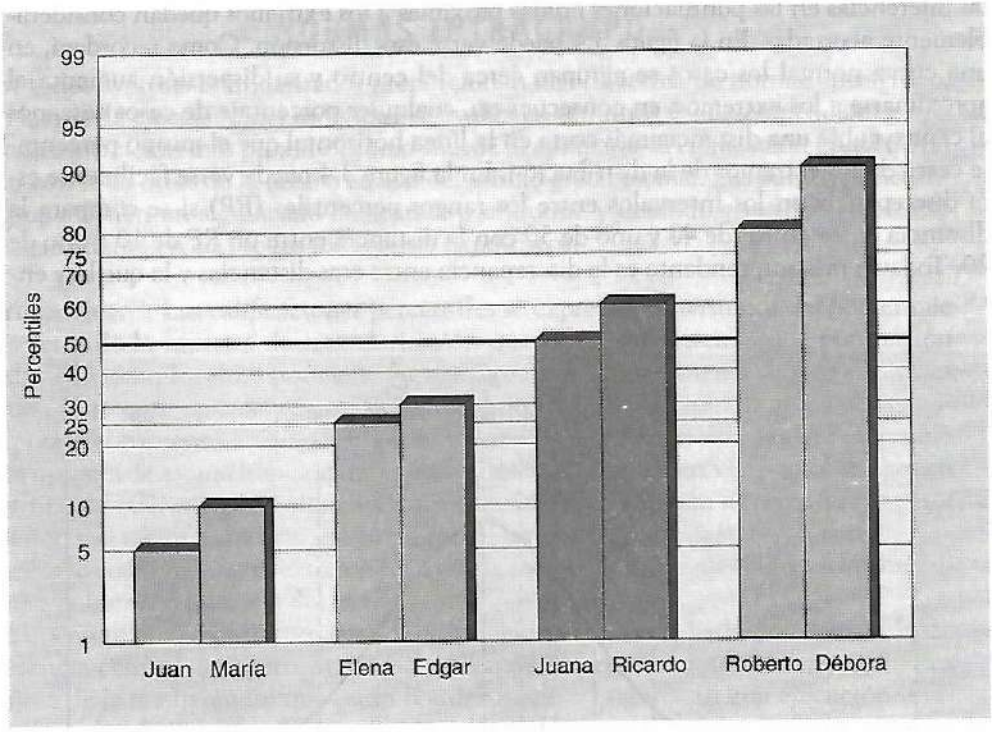


Figura 3.5. Diagrama de percentil normal. Los percentiles se espacian de modo que correspondan a distancias iguales en una distribución normal. Compare la distancia entre las puntuaciones obtenidas por Juan y María con la que existe entre los resultados de Elena y Edgar; dentro de ambos pares, la diferencia percentil es de 5 puntos. Juana y Ricardo difieren en 10 puntos percentiles, igual que Roberto y Débora.

quier caso, la diferencia real entre puntuaciones aparecerá representada correctamente. Muchas baterías de aptitudes y de aprovechamiento utilizan esta técnica en sus perfiles de calificaciones, que muestran el desempeño del individuo en cada prueba.

Calificaciones estándares. Los instrumentos actuales hacen un uso creciente de las calificaciones estándares, que desde cualquier punto de vista constituyen el tipo más satisfactorio de puntuación derivada. Estas calificaciones expresan la distancia del individuo de la media en términos de la desviación normal de la distribución normal.

Las calificaciones estándares pueden obtenerse por medio de transformaciones lineales o no lineales de las puntuaciones crudas originales. Las calificaciones estándares obtenidas por medio de una *transformación lineal* conservan las relaciones numéricas exactas de las puntuaciones originales gracias a que para calcularlas se resta una constante de cada puntuación cruda y se divide el resultado entre otra constante. La magnitud relativa de las diferencias entre las calificaciones estándares así derivadas corresponde exactamente a la que existe entre las puntuaciones crudas. Todas las propiedades de la distribución original de las puntuaciones crudas se duplican en la distribución de calificaciones estándares, lo cual implica que en las dos clases de puntuaciones es posible realizar el mismo cálculo sin distorsionar de los resultados.

Las puntuaciones estándares derivadas linealmente suelen conocerse simplemente como "calificaciones estándares" o como "calificaciones z". Para calcular una calificación z, se encuentra la diferencia entre la calificación individual cruda y la media del grupo normativo, y luego se divide la diferencia entre la *DE* del grupo normativo. La tabla 3.3 muestra el cálculo de las puntuaciones z para dos individuos, de los que uno cae a una *DE* por encima de la media del grupo y el otro a .40 *DE* por debajo. Cualquier puntuación no procesada que sea exactamente igual a la media es equivalente a una puntuación z de cero. Es evidente que dicho procedimiento producirá puntuaciones derivadas con signo negativo para todos los que caigan por debajo de la media. Además, como el rango total de la mayor parte de los grupos no se extiende más allá de tres *DE* por encima y por debajo de la media, las puntuaciones estándares deben expresarse con al menos un decimal para permitir la suficiente diferenciación.

Las dos condiciones mencionadas (la ocurrencia de valores negativos y de decimales) tienden a producir números desproporcionados que resultan confusos y difíciles

Tabla 3.3

Cálculo de puntuaciones estándares

$$z = \frac{X - M}{DE} \qquad M = 60 \qquad DE = 5$$

Puntuación de Elena	*Puntuación de Roberto*
$X_1 = 65$	$X_2 = 58$
$z_1 = \frac{65 - 60}{5}$	$z_2 = \frac{58 - 60}{5}$
$= +1.00$	$= -0.40$

de usar en cálculos o informes, por lo que suele aplicarse alguna otra transformación lineal para poner las puntuaciones en una forma más conveniente; por ejemplo, las puntuaciones de las Pruebas de Evaluación Académica (*Scholastic Assessment Tests*, SAT) de la Junta Universitaria son calificaciones estándares ajustadas a una media de 500 y una *DE* de 100. Así, en esta prueba una calificación estándar de –1 se expresa como 400 (500 – 100 = 400), mientras que una de +1.5 corresponde a 650 (500 + 1.5 × 100 = 650). Lo único que se requiere para convertir una calificación estándar original en la nueva escala es multiplicarla por la *DE* deseada (100) y sumar o restar el resultado de la media deseada (500). Es posible elegir arbitrariamente cualquier otra constante conveniente para las nuevas media y *DE*; por ejemplo, las puntuaciones obtenidas en los subtests separados de las Escalas de Inteligencia de Wechsler se convierten en una distribución con una media de 10 y una *DE* de 3. Todas esas medidas son ejemplos de calificaciones estándares transformadas en forma lineal.

Recordará que una de las razones para transformar las puntuaciones crudas en cualquier escala derivada es la necesidad de hacer comparables los resultados de diferentes pruebas. Las puntuaciones estándares derivadas linealmente sólo son comparables cuando provienen de distribuciones de más o menos la misma forma; por ejemplo, en tales condiciones, una calificación que corresponda a una *DE* por encima de la media significa que el individuo ocupa la misma posición en los dos grupos. En ambas distribuciones, la calificación supera aproximadamente al mismo porcentaje de sujetos, y este porcentaje puede ser determinado si se conoce la forma de la distribución; sin embargo, si una distribución es muy asimétrica y la otra es normal, una calificación *z* de +1.00 podría exceder sólo al 50 por ciento de los casos en un grupo y a más del 84 por ciento en el otro.

Cuando la forma de las distribuciones es distinta, las calificaciones pueden hacerse comparables empleando transformaciones no lineales que las ajuste a cualquier curva de distribución. La edad mental y las puntuaciones percentilares representan transformaciones no lineales, pero tienen otras limitaciones que ya vimos. Aunque en ciertas circunstancias puede ser más apropiada otra distribución, la curva normal suele emplearse con este propósito. Una de las razones para esta elección es que la mayor parte de las distribuciones de puntuaciones crudas se aproxima a la curva normal más que a cualquier otra. Además, las medidas físicas, como la estatura y el peso, que utilizan escalas de unidades iguales derivadas por operaciones físicas, en general producen distribuciones normales. Otra ventaja de la curva normal es que tiene muchas propiedades matemáticas útiles que facilitan los cálculos posteriores.

Las *calificaciones estándares normalizadas* son calificaciones estándares expresadas en términos de una distribución que ha sido transformada para ajustarse a una curva normal. Dichas calificaciones pueden calcularse recurriendo a tablas que muestran el porcentaje de casos que cae a diferentes distancias *DE* de la media de una curva normal. Para ello, primero se encuentra el porcentaje de individuos de la muestra de estandarización que cae en o por encima de cada puntuación cruda, este porcentaje se localiza luego en la tabla de frecuencia de la curva normal y se obtiene la correspondiente calificación estándar normalizada. Las calificaciones estándares normalizadas se expresan de la misma forma que las calificaciones estándares derivadas linealmente, es decir, con una media igual a cero y una *DE* igual a 1. Así, una calificación normalizada de cero indica que el individuo cae en la media de la curva normal, que supera al

Tabla 3.4

Porcentajes de la curva normal para su uso en la conversión estanina

Porcentaje	4	7	12	17	20	17	12	7	4
Estanina	1	2	3	4	5	6	7	8	9

50 por ciento del grupo. Una calificación de –1 significa que sobrepasa aproximadamente a 16 por ciento del grupo y una calificación de +1 que aventaja al 84 por ciento. Como puede verse en la línea inferior de la figura 3.4, esos porcentajes corresponden respectivamente a una distancia de una *DE* por debajo y una *DE* por encima de la media de una curva normal.

Igual que las calificaciones estándares derivadas linealmente, las calificaciones estándares normalizadas pueden adoptar cualquier forma conveniente. Si la calificación estándar normalizada se multiplica por 10 y se suma o resta de 50 se convierte en una *calificación* T, una puntuación que propuso primero W. A. McCall (1922) y en la que una puntuación de 50 corresponde a la media, una de 60 a una *DE* por encima de la media, etc. Otra transformación conocida es la de la escala *estanina*, elaborada por la aviación estadounidense durante la Segunda Guerra Mundial, que proporciona un sistema de calificaciones de un solo dígito con una media de 5 y una *DE* aproximadamente igual a 2.[4] El nombre, que en inglés (*stanine*) es una contracción de "*stan*dard *nine*" o "estándar nueve", se basa en el hecho de que las calificaciones van de 1 a 9. La restricción de las calificaciones a números de un solo dígito tiene ciertas ventajas para el cálculo, ya que cada calificación requiere una sola columna en la computadora.

Las puntuaciones crudas pueden convertirse fácilmente en estaninas si se disponen los resultados originales en orden de tamaño y se les asignan luego estaninas de acuerdo con los porcentajes de la curva normal reproducidos en la tabla 3.4; por ejemplo, si el grupo consta exactamente de 100 personas, las cuatro con la calificación más baja reciben una calificación estanina de 1, las siguientes siete una calificación de 2, las siguientes 12 una calificación de 3, etc. Cuando el grupo contiene más o menos 100 casos, primero se calcula el número correspondiente a cada porcentaje designado y luego recibe las estaninas apropiadas. Así, si son 200 casos, a ocho les sería asignada una estanina de 1 (4 por ciento de 200 = 8), con 150 casos seis recibirían una estanina de 1 (4 por ciento de 150 = 6). Para cualquier grupo que contenga de 10 a 100 casos, Bartlett y Edgerton (1966) prepararon una tabla que permite convertir directamente los rangos en estaninas. Debido a sus ventajas prácticas y teóricas, las estaninas han sido empleadas ampliamente, en especial con las pruebas de aptitud y de aprovechamiento.

Aunque las calificaciones estándares normalizadas son el tipo de puntuación más adecuado para la mayor parte de los propósitos, hay ciertas objeciones técnicas a la

[4] Kaiser (1958) propuso una modificación de la escala estanina que hace ligeros cambios en los porcentajes y produce una *DE* exactamente igual a 2, lo que facilita su manejo cuantitativo. Otras variantes son la escala C (Guilford y Fruchter, 1978, pp. 484–487), que consta de 11 unidades y también genera una *DE* de 2, y la escala *sten* de 10 unidades con cinco unidades por encima y cinco por debajo de la media (Canfield, 1951).

normalización rutinaria de todas las distribuciones. Dicha transformación sólo debería llevarse a cabo cuando la muestra es grande y representativa y cuando existen razones para creer que la desviación de la normalidad resulta de defectos en la prueba más que de las características de la muestra y otros factores que afecten la conducta considerada. También debe observarse que cuando la distribución original de puntuaciones crudas se aproxima a la normalidad, las calificaciones estándares derivadas linealmente y las calificaciones estándares normalizadas resultan muy similares a pesar de las diferencias entre los métodos empleados para derivarlas. Como es obvio, el proceso de normalizar una distribución que de hecho es prácticamente normal producirá poco o ningún cambio. Siempre que sea factible suele ser más deseable obtener una distribución normal de puntuaciones crudas por el ajuste adecuado del nivel de dificultad de los reactivos de la prueba más que por la normalización subsecuente de una distribución marcadamente no normal. Con una distribución más o menos normal de puntuaciones crudas, las calificaciones estándares linealmente derivadas cumplirán el mismo propósito que las calificaciones estándares normalizadas.

El CI de desviación. El CI (cociente de inteligencia) se introdujo en los primeros tests de inteligencia en un esfuerzo por convertir las calificaciones de EM en un indicador uniforme de la posición relativa del individuo. El CI era simplemente la razón de la edad mental y la edad cronológica multiplicada por 100 para eliminar los decimales (CI = 100 × EM/EC). Por supuesto, si la EM de una niña es igual a su EC su CI sería exactamente igual a 100. Un CI de 100 representa entonces la ejecución normal o promedio, en tanto que uno por debajo indica "retardo" y otro por encima "aceleración".

Sin embargo, la aparente simplicidad lógica del CI tradicional pronto decepcionó. Un problema técnico importante es que, a menos que la *DE* de la distribución de CI permanezca aproximadamente constante con la edad, los CI correspondientes a diferentes niveles de edad no resultan comparables; por ejemplo, un CI de 115 a los 10 años puede indicar el mismo grado de superioridad que uno de 125 a los 12, porque ambos pueden caer a una *DE* de la media de su respectiva distribución de edad. En la práctica, ha resultado muy difícil establecer pruebas que satisfagan los requisitos psicométricos para hacer comparables a los CI en todo el rango de edad. Es principalmente por eso que la razón de CI ha sido reemplazado por el llamado CI de desviación, que es otra variante de las calificaciones estándares. El CI de desviación es una calificación estándar con una media de 100 y una *DE* similar a la utilizada en la distribución de CI del Stanford-Binet. Aunque la *DE* de la razón de CI del Stanford-Binet (utilizada por última vez en la edición de 1937) no permanece exactamente constante en todas las edades, fluctúa alrededor de una media ligeramente mayor a 16, de ahí que si se elige una *DE* de 16 al informar las calificaciones estándares de una prueba de desarrollo reciente, las puntuaciones pueden ser interpretadas de la misma manera que la razón de CI del Stanford-Binet. Como el CI del Stanford-Binet ha estado en uso durante mucho tiempo, examinadores y clínicos se han acostumbrado a interpretar y clasificar las ejecuciones en dicha prueba en términos de los niveles de CI. Han aprendido lo que pueden esperar de individuos con CI de 40, 70, 90, 130, etc., por lo que hay ciertas ventajas prácticas en el uso de una escala que corresponda a las conocidas distribuciones de CI del Stanford-Binet. Dicha correspondencia de unidades de calificación se consigue en la selección de las cifras de la media y la *DE* que concuerden con los de la distribución de Stanford-Binet.

Debe agregarse que el uso del término "CI" para designar a esas calificaciones estándares puede resultar engañosa. Estos CI no se obtienen a partir de los mismos métodos que se emplean para encontrar las razones CI tradicionales; no son razones de edades mentales y edades cronológicas. La justificación para su uso estriba en la familiaridad general del término "CI" y en el hecho de que estas calificaciones pueden interpretarse como CI en tanto que su *DE* sea aproximadamente igual a la de los CI ya conocidos. Entre las primeras pruebas en expresar las calificaciones en términos de CI de desviación están las Escalas de Inteligencia de Wechsler, cuya media es de 100 y la *DE* es igual a 15. El CI de desviación también se emplea en varias pruebas colectivas de inteligencia y en la tercera revisión del Stanford-Binet (1960).

Con el uso creciente del CI de desviación se hace importante recordar que los CI de desviación de diferentes pruebas sólo son comparables si emplean valores iguales o muy similares para la *DE*, que además tienen que aparecer siempre en el manual y el usuario de la prueba debe considerarlas con atención. Si el autor elige valores diferentes para la *DE* al elaborar la escala del CI de desviación, el significado de cualquier CI en tal prueba será muy diferente del que tenga en otras. Esas discrepancias se ilustran en la tabla 3.5, que muestra el porcentaje de casos que obtendrían diferentes niveles de CI en distribuciones normales con *DE* que varían entre 12 y 18, valores que por cierto han sido realmente utilizados en las escalas CI de pruebas publicadas; por ejemplo, la tabla 3.5 muestra que un CI de 70 corta el 3.1 por ciento inferior cuando la *DE* es igual a 16 (como el Stanford-Binet), pero que puede separar apenas el 0.7 por ciento (*DE* = 12) o hasta el 5.1 (*DE* = 18). Por lo común, se ha empleado el CI de 70 como punto de corte para identificar el retardo mental clínicamente significativo. Por supuesto, las mismas discrepancias se aplican a los CI de 130 y superiores, que pueden usarse al seleccionar a niños para programas especiales para sobredotados. Dependiendo de la prueba elegida, el CI entre 90 y 110, descrito como el normal, puede incluir tan poco como el 42 o tanto

Tabla 3.5

Porcentaje de casos de cada intervalo de CI en las distribuciones normales con media de 100 y diferentes desviaciones estándares

Intervalo CI	Porcentaje de frecuencias			
	DE = 12	DE = 14	DE= 16	DE = 18
130 y por encima	0.7	1.6	3.1	5.1
120–129	4.3	6.3	7.5	8.5
110–119	15.2	16.0	15.8	15.4
100–109	29.8 } 59.6	26.1 } 52.2	23.6 } 47.2	21.0 } 42.0
90– 99	29.8 }	26.1 }	23.6 }	21.0 }
80– 89	15.2	16.0	15.8	15.4
70– 79	4.3	6.3	7.5	8.5
Debajo de 70	0.7	1.6	3.1	5.1
Total	100.0	100.0	100.0	100.0

(Cortesía de The Psychological Corporation.)

como el 59.6 por ciento de la población. Para estar seguros, los editores de las pruebas se esfuerzan por adoptar de manera uniforme una *DE* de 16 tanto en las nuevas pruebas como en las nuevas ediciones de las anteriores; sin embargo, las variaciones aún son suficientes como para hacer que la supervisión de la *DE* resulte imperativa.

Interrelaciones de las calificaciones intragrupo. En esta etapa de la revisión de las calificaciones derivadas, el lector puede haber advertido la similitud entre las distintas calificaciones. Los percentiles han ido adquiriendo gradualmente al menos un parecido gráfico con las calificaciones estándares normalizadas. Si la distribución original de las puntuaciones no procesadas se aproxima a la curva normal resulta difícil distinguir a las calificaciones estándares lineales de las calificaciones estándares normalizadas. Por último, las calificaciones estándares se han convertido en CI y viceversa. En relación con ese último punto, un nuevo examen del significado original del CI en tests como el Stanford-Binet demostrará que es posible interpretarlo como calificaciones estándares. Si sabemos que la distribución de los CI del Stanford-Binet tiene una media de 100 y una *DE* aproximadamente igual a 16, podemos concluir que un CI de 116 cae a una distancia de una *DE* por encima de la media y representa una calificación estándar de +1.00. De modo similar, un CI de 132 corresponde a una calificación estándar de +2.00, un CI de 76 a una calificación estándar de −1.50, etc. Más aún, una razón de CI del Stanford-Binet igual a 116 corresponde a un percentil de aproximadamente 84 porque en una curva normal 84 por ciento de los casos cae por debajo de +1.00 *DE* (figura 3.4).

La figura 3.6 resume las relaciones en una distribución normal entre las calificaciones que hemos estudiado hasta ahora. Entre ellas se incluyen las calificaciones *z*, las calificaciones del Examen de Admisión a la Universidad (*College Entrance Examination Board*, CEEB), el CI de desviación de Wechsler (*DE* = 15), calificaciones *T*, estaninas y percentiles. Los CI de cualquier prueba han de coincidir con la escala dada del CI de desviación siempre que se distribuyan normalmente y tengan una *DE* igual a 15. Cualquier otro CI normalmente distribuido puede agregarse a la tabla si conocemos su *DE*; por ejemplo, si la *DE* es de 20, entonces un CI de 120 corresponde a +1 *DE*, un CI de 80 a −1 *DE*, etcétera.

En conclusión, la forma exacta en que se informan los resultados se determina sobre todo por la conveniencia, familiaridad o facilidad para establecer las normas. Las calificaciones estándares de cualquier forma (incluyendo el CI de desviación) han llegado a reemplazar a otras calificaciones por las ventajas que ofrecen en relación con la elaboración de la prueba y el tratamiento estadístico de los datos; no obstante, la mayor parte de las calificaciones derivadas intragrupo son en esencia similares si se transforman con cuidado y se interpretan adecuadamente. Cuando se satisfacen ciertas condiciones estadísticas, todas pueden ser traducidas con facilidad a cualquiera de las otras.

RELATIVIDAD DE LAS NORMAS

Comparaciones entre pruebas. El CI, o cualquier otra calificación, debe acompañarla siempre el nombre del test del que se obtuvo. Los resultados de las pruebas no pueden interpretarse en abstracto, tienen que ser referidos a pruebas particulares. Si los registros escolares muestran que Pedro Gómez obtuvo un CI de 94 y Teresa Olivo un CI de 110, estas calificaciones no pueden aceptarse porque aparecen sin contar con

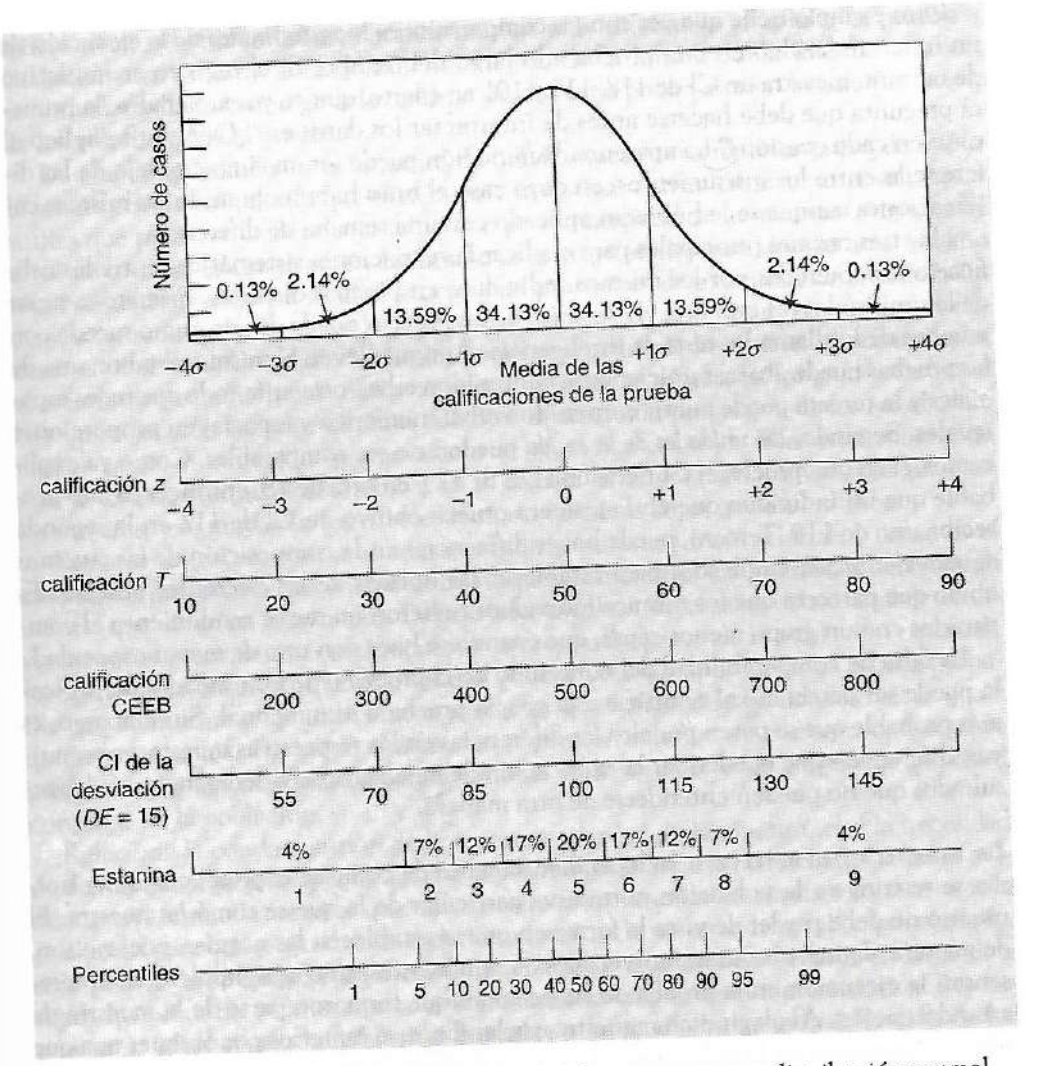


Figura 3.6. Relaciones entre diferentes calificaciones en una distribución normal.

mayor información. Las posiciones de esos dos estudiantes podría invertirse al cambiar las pruebas que cada uno presentó en su respectiva escuela.

Del mismo modo, la posición relativa de un individuo en diferentes funciones puede ser muy mal representada por la falta de comparabilidad de las normas de una prueba. Supongamos que a una estudiante se le aplica una prueba de comprensión verbal y un test de aptitud espacial para determinar su ubicación relativa en ambos campos. Si la prueba de habilidad verbal se estandarizó con una muestra aleatoria de estudiantes de secundaria mientras que la prueba espacial se estandarizó en un grupo selecto de asistentes a cursos para elección de carrera, el examinador podría llegar a la conclusión errónea de que la capacidad verbal de la chica es mucho mayor que la espacial, cuando podría ser el caso contrario.

Otro ejemplo tiene que ver con las comparaciones longitudinales de la ejecución de un único individuo en una prueba a lo largo del tiempo. Si el registro acumulativo de un niño muestra un CI de 118, 115 y 101 en cuarto, quinto y sexto grados, la primera pregunta que debe hacerse antes de interpretar los datos es: "¿Qué test se aplicó al niño en cada ocasión?" La aparente disminución puede ser un simple reflejo de las diferencias entre los instrumentos, en cuyo caso el niño habría obtenido las mismas calificaciones aunque se le hubieran aplicado con una semana de diferencia.

Hay tres razones principales para explicar las variaciones sistemáticas entre las calificaciones obtenidas por los mismos individuos en pruebas distintas. Primero, a pesar de la similitud en el nombre, el *contenido* de las pruebas puede diferir, como sucede con muchos de los llamados tests de inteligencia. Aunque lleven el mismo nombre, una de las pruebas puede abarcar únicamente contenido verbal, otra ante todo aptitudes espaciales y la tercera puede cubrir contenido verbal, numérico y espacial en proporciones iguales. Segundo, las *unidades de la escala* pueden no ser comparables. Como ya explicamos, si en una prueba el CI tiene una *DE* de 12 y en otra de 18, entonces es más probable que un individuo que en la primera prueba obtuvo un CI de 112 en la segunda reciba uno de 118. Tercero, puede haber diferencias en la composición de las *muestras de estandarización* empleadas para establecer las normas de las diferentes pruebas. Es obvio que parecerá que los mismos individuos tuvieron un mejor rendimiento al compararlos con un grupo menos capaz, que cuando se hace con uno de mayor capacidad.

La falta de comparabilidad del contenido de la prueba o de las unidades de la escala puede ser detectado al remitirse a la propia prueba o a su manual. Sin embargo, es más probable que se pasen por alto las diferencias en las respectivas muestras normativas, diferencias que pueden ser la explicación de muchas otras discrepancias en los resultados que no pueden entenderse de otra manera.

La muestra normativa. Independientemente de como se exprese, cualquier norma se restringe a la población normativa particular de la que se tomó la muestra. El usuario no debe perder de vista la forma en que se establecen las normas, pues no son, de manera alguna, absolutas, universales o permanentes, sino que simplemente representan la ejecución en la prueba de las personas que formaron parte de la muestra de estandarización. Al elegir dicha muestra, suelen hacerse esfuerzos por obtener una que sea representativa de la población a la que va dirigida la prueba.

En la terminología estadística se distingue entre *muestra* y *población*. La primera se refiere al grupo de individuos realmente examinado, mientras que la población designa al grupo mayor, pero de similar constitución, de donde se extrajo la muestra; por ejemplo, si desea establecer las normas de ejecución en una prueba para la población de niños varones de 10 años que asisten a escuelas públicas urbanas, puede aplicar el instrumento a una muestra cuidadosamente elegida de 500 muchachos de 10 años que asisten a escuelas públicas en varias ciudades del país. Hay que verificar la muestra en lo que atañe a la distribución geográfica, el nivel socioeconómico, la composición étnica y otras características pertinentes para asegurar que representa de verdad a la población definida.

En el desarrollo y la aplicación de las normas debe dedicarse mucha atención a la muestra de estandarización. Es evidente que la muestra en la que se basan las normas debe ser lo suficientemente grande como para proporcionar valores estables, lo que

quiere decir que no se deben obtener normas muy diferentes al aplicar la prueba a cualquier otra muestra cuidadosamente elegida de la misma población. Las normas con un error de muestreo grande tendrán desde luego poco valor en la interpretación de los resultados de la prueba.

De igual importancia es el requisito de que la muestra sea representativa de la población considerada, por lo que es necesario investigar con cuidado factores selectivos sutiles que pudieran haber disminuido la representatividad de la muestra. Algunos de estos factores se encuentran en las muestras institucionales que, debido a su tamaño relativamente grande y a su fácil acceso para propósitos de examinación, resultan un campo tentador para la acumulación de datos normativos. De cualquier forma, es necesario analizar con cuidado las limitaciones especiales de las muestras; por ejemplo, al examinar a estudiantes se observará que conforme se avanza en los grados escolares hay una selección superior de casos debido a la deserción progresiva de los estudiantes menos aptos, eliminación que no afecta de igual manera a los diferentes subgrupos. Así, la tasa de eliminación selectiva en la escuela es mayor para los hombres que para las mujeres, y es mayor en los niveles socioeconómicos bajos que en los altos.

Factores selectivos similares operan en otras muestras institucionales, como es el caso de los prisioneros, los pacientes de hospitales mentales o las personas con retardo mental que viven en internados. Debido a los muchos factores especiales que determina la propia internación, esos grupos no son representativos de la población entera de los infractores, los trastornados o los que padecen retardo; por ejemplo, es más probable que se interne a los individuos retardados que sufren discapacidades físicas que a quienes son aptos. De modo similar, la proporción relativa de personas con retardo profundo será mayor en las muestras tomadas en instituciones de asistencia que en el conjunto de la población.

Relacionada con la cuestión de la representatividad de la muestra, está la necesidad de definir la población específica a la que pueden generalizarse las normas. Como es evidente, una forma de asegurar que una muestra es representativa consiste en restringir la población para que se ajuste a las especificaciones de la muestra disponible; por ejemplo, si la población se define para incluir únicamente a escolares de 14 años en lugar de a todos los niños de 14 años, entonces una muestra escolar sería representativa. Por supuesto, lo ideal es que la población deseada se defina de antemano en función de los objetivos de la prueba, para formar luego una muestra adecuada. No obstante, los obstáculos prácticos para obtener participantes puede hacer que esta meta resulte inalcanzable. En dicho caso, es mucho mejor redefinir la población de manera más reducida que señalar normas sobre una población ideal que no está adecuadamente representada en la muestra de estandarización. En la práctica real, son muy pocas las pruebas que se estandarizan sobre poblaciones tan amplias como suele suponerse, y ¡ninguna prueba ofrece normas para la especie humana! Además, es dudoso que muchas pruebas brinden normas verdaderamente adecuadas para poblaciones tan amplias como "los europeos adultos", "los niños mexicanos de 10 años", etc. En consecuencia, las muestras obtenidas por diferentes autores de pruebas pueden no ser representativas de las poblaciones alegadas y presentar diversos sesgos, de ahí que las normas resultantes no sean comparables.

Al interpretar los resultados de la prueba, el usuario debe tomar en consideración las influencias específicas que pueden haber actuado sobre la muestra normativa utili-

zada al estandarizarla, y que incluyen factores selectivos especiales así como las condiciones sociales prevalecientes en el momento en que fueron obtenidos los datos normativos (Anastasi, 1985b).

Normas fijas nacionales. Una solución a la falta de comparabilidad de las normas consiste en usar una prueba ancla para desarrollar tablas de equivalencia para los resultados de distintas pruebas (para mostrar qué puntuación de la prueba A es equivalente a cada puntuación de la prueba B). Esto puede lograrse con el *método equipercentil*, en el que las puntuaciones se consideran equivalentes cuando tienen percentiles iguales en un determinado grupo; por ejemplo, si en el mismo grupo el percentil 80 corresponde a un CI de 115 en la prueba A y a un CI de 120 en la prueba B, entonces el CI de 115 de la prueba A se considera equivalente al CI de 120 de la prueba B. Este procedimiento ha sido adoptado de manera restringida por algunos editores en la preparación de tablas de equivalencias para algunos de sus instrumentos (por ejemplo, Lennon, 1966a).

Ocasionalmente se hacen propuestas más ambiciosas para calibrar cada nueva prueba contra una sola prueba ancla que haya sido administrada a una muestra normativa representativa a nivel nacional (Lennon, 1966b). Un ejemplo de este procedimiento se encuentra en el Estudio de Prueba Ancla (*Anchor Test Study*) conducido por el Servicio de Pruebas Educativas (*Educational Testing Service*) bajo los auspicios de la Oficina de Educación de los Estados Unidos (Jaeger, 1973). Dicho estudio representa un esfuerzo sistemático por proporcionar normas comparables y verdaderamente representativas de la población estadounidense para siete de las pruebas de aprovechamiento en lectura aplicadas a niños de educación elemental. Gracias a un diseño experimental bien controlado, en 50 estados de ese país se examinó a más de 300 000 escolares de cuarto, quinto y sexto grado. La prueba ancla estaba formada por las subpruebas de lectura de comprensión y de vocabulario de la Prueba de Aprovechamiento Metropolitana (*Metropolitan Achievement Test*), para la cual se establecieron nuevas normas en una fase del proyecto. En la fase de calibración del estudio cada niño presentó las subpruebas de lectura de comprensión y vocabulario de dos de las siete baterías, y cada una se apareó a la vez con todas las demás. Algunos grupos recibieron formas paralelas de las dos subpruebas de la misma batería; en otros grupos todos los pareamientos fueron duplicados en una secuencia inversa para controlar el orden de la administración. A partir de los análisis estadísticos de todos esos datos se utilizó el método equipercentil para preparar tablas de equivalencia de puntuaciones para las siete pruebas. También se preparó un manual de interpretación de los resultados para uso de los sistemas escolares y otros interesados (Loret, Seder, Bianchini y Vale, 1974).

Los datos de la fase de calibración del Estudio de Prueba Ancla se utilizaron después para formar una sola escala de puntuación, conocida como la Escala Nacional de Referencia (*National Reference Scale*, Rentz y Bashaw, 1977). La tabla de conversión permite la transformación de una puntuación de cualquier forma de las siete pruebas en cualquiera de los niveles escolares en una calificación de tres lugares sobre una escala uniforme y continua. Esta escala fue elaborada empleando el análisis de reactivos y los métodos de escalamiento del modelo Rasch, uno de los modelos de análisis de reactivos más sencillos del que veremos más aquí y en el capítulo 7.

Para muchos propósitos de examinación resulta útil contar con puntuaciones comparables de diferentes pruebas que se expresen sobre una escala uniforme de medición

y se refieran a una sola muestra normativa. Empero, debe observarse que existen diferentes grados y clases de comparabilidad de las calificaciones. La comparabilidad en una situación particular depende de la similitud de las pruebas en cuanto al contenido, a propiedades psicométricas como la confiabilidad y el grado de dificultad y a los procedimientos estadísticos utilizados para lograr la comparabilidad (Angoff, 1984; Angoff y Cowell, 1986; P. W. Holland y Rubin, 1982). Las pruebas no deben considerarse igualadas o equivalentes a menos que sean verdaderamente intercambiables. No obstante, diferentes clases y grados de comparabilidad *pueden* facilitar la interpretación de los resultados siempre que las puntuaciones comparables se usen de modo apropiado y con pleno conocimiento de la forma en que se derivaron.

Normas específicas. Otro acercamiento a la no equivalencia de las normas existentes —que quizá sea el más realista para la mayor parte de los instrumentos— consiste en estandarizarlas sobre poblaciones definidas de manera más restringida y elegidas para adaptarse a los propósitos específicos de cada prueba. En esos casos, los límites de la población normativa debe informarse claramente con las normas. Así, puede decirse que las normas se aplican a "los empleados administrativos de las grandes empresas" o a "los estudiantes de primer año de ingeniería". Para muchos propósitos de examinación resulta deseable contar con normas muy específicas. Aunque pueda disponerse de normas representativas de una población más amplia, siempre que subgrupos identificables obtengan puntuaciones apreciablemente diferentes en una prueba particular resulta útil informar por separado las *normas del subgrupo*. Los subgrupos pueden formarse de acuerdo con la edad, el grado escolar, el programa de estudios, el sexo, la región geográfica, el ambiente urbano o rural, el nivel socioeconómico y muchas otras variables. El uso que vaya a hacerse de la prueba determina la diferenciación que resulte más pertinente, así como la conveniencia de contar con normas generales o específicas.

Debemos mencionar también el caso de las *normas locales*, que con frecuencia son establecidas por los propios usuarios en ambientes particulares. Los grupos empleados para derivar dichas normas se definen de manera aún más limitada que los subgrupos considerados antes. Así, un empleador puede acumular normas de los solicitantes para un determinado trabajo en una empresa en particular, o la oficina de admisión de una universidad puede evaluar el rendimiento de los alumnos en términos de su propia distribución de calificaciones. Esas normas locales son más apropiadas que las normas nacionales para muchos propósitos de examinación, como la predicción del desempeño en el empleo o académico, la comparación del aprovechamiento relativo de un niño en diferentes materias o bien la medición del progreso de un individuo a lo largo del tiempo.

Grupo fijo de referencia. Aunque la forma en que se calcula la mayor parte de las calificaciones derivadas permite una interpretación normativa inmediata de la ejecución en una prueba, hay algunas excepciones notables. Una escala no normativa utiliza un grupo fijo de referencia para asegurar la *comparabilidad* y *continuidad* de las calificaciones sin brindar una evaluación normativa del desempeño. Con dicha escala, la interpretación normativa tiene que referirse a normas obtenidas de manera independiente de una población adecuada. Con este propósito suelen emplearse normas locales u otras normas específicas.

Uno de los primeros ejemplos de escalamiento en términos de un grupo fijo de referencia se encuentra en la escala de calificación del Test de Aptitud Académica de la Junta Universitaria[5] (*College Board Scholastic Aptitude Test*, Donlon, 1984). Entre 1926 (cuando se aplicó la prueba por primera vez) y 1941, las calificaciones del SAT se expresaban sobre una escala normativa en términos de la media y la *DE* de los candidatos que presentaban la prueba en cada aplicación. Dado el incremento en el número y la variedad de las universidades participantes en la Junta Universitaria y los cambios en la composición de la población de aspirantes, se concluyó que era necesario mantener la continuidad de la escala. De otra manera, el resultado de un individuo dependería de las características del grupo examinado en un año particular. Un motivo aún más urgente para mantener la continuidad de la escala proviene de la observación de que, debido a la operación diferencial de factores selectivos, los estudiantes que presentaban el SAT en cierta época del año obtenían un resultado inferior al obtenido por quienes lo presentaban en otros momentos. Por ende, después de 1941 todas las calificaciones del SAT se expresaron en términos de la media y la *DE* de los aproximadamente 11 000 candidatos que presentaron la prueba en 1941. Esos aspirantes constituyeron el grupo fijo de referencia empleado en el escalamiento de las formas subsecuentes del instrumento. Así, una puntuación de 500 en cualquier forma del SAT corresponde a la media de la muestra de 1941, una calificación de 600 cae a una *DE* por encima de la media, etcétera.

Para permitir la transformación de puntuaciones crudas de cualquier forma del SAT a esas calificaciones de un grupo fijo de referencia en cada forma se incluyó una pequeña prueba ancla (o conjunto de reactivos comunes). En consecuencia, cada nueva forma se ligaba a una o dos pruebas anteriores, las que a su vez estaban ligadas a otras formas por una cadena de reactivos que se extendían hasta la forma de 1941. Estas calificaciones del SAT no normativas podían entonces interpretarse al compararlas con cualquier distribución apropiada de puntuaciones, como las de una universidad particular, una clase de universidad, una región, etc. Esas normas específicas son más útiles al tomar decisiones de admisión a la universidad que las normas anuales basadas en toda la población de candidatos. Más aún, cualquier cambio en la población de aspirantes puede ser detectada con una única escala fija de calificaciones. Más recientemente, la escala SAT ha vuelto a centrarse en el desempeño de más de un millón de estudiantes estadounidenses que se graduaron de secundaria en 1990 y presentaron el test en su penúltimo o último año escolar. Los resultados para los estudiantes que presentaron el SAT después del primero de abril de 1995 se informaron de acuerdo con la escala "reciente" derivada del grupo de referencia de 1990. Se han elaborado materiales auxiliares e interpretativos para ayudar a los usuarios a convertir las calificaciones individuales y agregadas de la escala anterior y viceversa (véase el capítulo 17), lo que permite una interpretación completa y diversificada de la ejecución individual para propósitos específicos de examinación.[6]

Las escalas formadas a partir de un grupo fijo de referencia son análogas en un aspecto a las escalas empleadas en la medición física. A este respecto, Angoff (1962, pp. 32–33) comentaba:

---

[5] Esta prueba recibió luego el nombre de Test de Evaluación Escolar (*Scholastic Assessment Test, SAT*) para mostrar el cambio de la orientación hacia la naturaleza de las calificaciones de las pruebas que surgió a finales del siglo XX (véase especialmente el capítulo 12 para una revisión del efecto de las diferencias de las experiencias personales sobre la ejecución en la prueba).

[6] Reconocemos la ayuda de Wayne Camara de la Junta Universitaria para la obtención de la información.

> Difícilmente existe una persona que conozca la definición original precisa de la longitud del pie usado en la medición de estatura o distancia, o que sepa quién fue el rey cuyo pie se acordó originalmente que fuera la norma; por otro lado, no hay nadie aquí que no sepa cómo evaluar longitudes y distancias en términos de esta unidad. Nuestra ignorancia del significado original de pie de manera alguna disminuye la utilidad que tiene para nosotros. Su utilidad procede del hecho de que se mantiene constante y nos permite familiarizarnos con él. No es necesario decir que las mismas consideraciones se aplican precisamente a otras unidades de medición —la pulgada, la milla, los grados Fahrenheit—. Del mismo modo, en el campo de la medición psicológica es razonable decir que la definición original de la escala no debe ser de importancia. Lo que es importante es mantener una escala constante —la cual, en el caso de un programa de examinación de formas múltiples, se consigue mediante una rigurosa igualación entre formas— y la disposición de los datos normativos complementarios para ayudar en la interpretación y la toma de decisiones, datos que deben revisarse de vez en cuando según lo justifiquen las circunstancias.

Teoría de las respuestas a los ítemes. Desde los setenta se observa un resurgimiento del interés en una clase de procedimientos de gran complejidad matemática para escalar la dificultad de los reactivos o ítemes de una prueba (Hambleton, 1989; Hambleton, Swaminathan y Rogers, 1991; Jaeger, 1977). Debido a la gran cantidad de cálculos que requieren, sólo fueron viables cuando aumentó la disponibilidad de computadoras de alta velocidad. Aunque su complejidad y sus procedimientos matemáticos difieren, estas aproximaciones se agruparon originalmente con el título general de *modelos de rasgos latentes*. La medida básica que utilizan es la probabilidad de que un sujeto con determinada habilidad (el llamado rasgo latente) acierte en un reactivo de dificultad conocida; sin embargo, ello no implica que existan rasgos latentes o habilidades subyacentes en un sentido físico o fisiológico, ni tampoco que originen una conducta. Los rasgos latentes son constructos estadísticos derivados matemáticamente de relaciones empíricas observadas entre las respuestas a la prueba. Una estimación inicial del rasgo latente de un examinado es la calificación total que obtiene en la prueba. Para evitar la falsa impresión creada por el término "rasgo latente", algunos de los principales exponentes de estos modelos lo han sustituido por el término más descriptivo y preciso de "Teoría de Respuesta al Ítem" o TRI (Lord, 1980; D. J. Weiss y Davison, 1981). Esta designación ha alcanzado un uso general dentro de la psicología.

En esencia, los modelos TRI se han usado para establecer una escala de medición uniforme y "libre de muestra" que pueda aplicarse a individuos y grupos de muy diversos niveles de habilidad y que permita probar contenidos de distintos niveles de dificultad. Como en el caso del grupo fijo de referencia que explicamos en la sección anterior, los modelos TRI requieren de reactivos ancla o una prueba común como puente entre las muestras examinadas y las pruebas o conjuntos de reactivos; sin embargo, en lugar de usar la media y *DE* de un grupo específico de referencia para definir el origen y tamaño de las unidades de la escala, los modelos TRI los establecen en términos de los datos que representan una amplia gama de habilidades y dificultades de los reactivos, que pueden provenir de varias muestras. Por lo general, el origen se establece cerca del centro de este rango. La unidad común de la escala se deriva matemáticamente de los datos de los reactivos, lo que tiene varias ventajas teóricas y prácticas

sobre los anteriores procedimientos de análisis de reactivos. Revisaremos después la metodología específica, en el capítulo 7, cuando tratemos las técnicas de análisis de reactivos. La TRI se ha ido incorporando en los programas de examinación a gran escala; por ejemplo, desde 1982 el procedimiento se emplea para igualar las calificaciones totales de las nuevas formas del SAT y expresarlas en una escala continua uniforme (Camara, Freeman y Everson, 1996; Donlon, 1984).

Aunque ha crecido la atención dirigida al problema general de *igualación de la prueba*, que permite expresar los resultados de diferentes formas de la misma prueba en una escala de calificación uniforme, las dificultades técnicas de los diversos planteamientos explorados con este propósito escapan al alcance de este libro, por lo que referimos al lector a P. W. Holland y Rubin (1982) y a Petersen, Kolen y Hoover (1989) para una revisión completa y una evaluación crítica de dicha metodología.

LAS COMPUTADORAS Y LA INTERPRETACIÓN DE LAS CALIFICACIONES DE LAS PRUEBAS

Desarrollos técnicos. Las computadoras han tenido un impacto importante en cada fase de la examinación, desde la elaboración de la prueba hasta la aplicación, calificación, interpretación y elaboración del reporte (F. B. Baker, 1989; Butcher, 1987; Gutkin y Wise, 1991; Roid, 1986). Los usos evidentes de las computadoras —y los encontrados antes— representaron un incremento sin precedentes en la rapidez con que pueden realizarse los procesos de calificación y análisis de datos. El uso de las computadoras en la aplicación automatizada de las pruebas convencionales también puede considerarse en esta categoría, en la medida en que proporciona formas más sencillas de aplicación. Empero, mucho más significativa es la contribución de las computadoras a la exploración de nuevos procedimientos y acercamientos a las pruebas psicológicas que habrían sido imposibles sin la flexibilidad y capacidad de procesamiento de datos que proporcionan. Este efecto de las computadoras lo ilustra la adopción creciente de los modelos TRI para el escalamiento libre de muestra. Veremos otras innovaciones del uso de las computadoras al revisar los temas pertinentes a lo largo del libro.

Así, examinaremos algunas aplicaciones de las computadoras a la evaluación del desempeño en las pruebas (F. B. Baker, 1989; Gutkin y Wise, 1991; Roid y Gorsuch, 1984). Al nivel más simple, casi todas las pruebas actuales, especialmente las diseñadas para aplicación colectiva, ahora están adaptadas para su *calificación computarizada*. Varias editoriales y organizaciones independientes de calificación de pruebas están equipadas para brindar esos servicios a los usuarios. Además, también se observa un aumento en el acceso a discos de computadora que los usuarios pueden emplear para calificar las pruebas en sus propias computadoras (por ejemplo, los programas ASSIST desarrollados por el Servicio Estadounidense de Asesoría). A un nivel más complejo, para algunas pruebas se dispone de *interpretaciones computarizadas* de los resultados, en las que el programa asocia informes verbales preparados con ciertos patrones de respuesta. Este sistema se ha empleado en test de aptitud y de personalidad; por ejemplo, en el Inventario Multifásico de Personalidad de Minnesota (*Minnesota Multiphasic Personality Inventory*, MMPI), que estudiaremos en el capítulo 13, los usuarios pueden obtener junto con las puntuaciones numéricas impresiones del diagnóstico y el informe interpretativo acerca de las tendencias de personalidad y la condición emocional

del examinado. Los usuarios que tienen acceso a sus propias computadoras pueden adquirir programas que además de las calificaciones numéricas también proporcionan informes interpretativos para ciertos tests, como las Escalas de Inteligencia para niños de Wechsler, Revisadas (WISC-R) y para adultos (WAIS-R).

Los *sistemas interactivos de cómputo* ilustran un nivel más complejo de interpretación individualizada de las puntuaciones de los tests. En éstos, el individuo está en contacto directo con la computadora por medio de estaciones de respuesta y entabla de hecho un diálogo con la computadora (J. A. Harris, 1973; Holtzman, 1970; M. R. Katz, 1974; Super *et al.*, 1970). Esta técnica ha sido investigada en relación con la planeación educativa y profesional y en la toma de decisiones. En estas situaciones, los resultados suelen incorporarse a la base de datos de la computadora junto con la información proporcionada por el estudiante o el cliente. En esencia, la computadora combina toda la información disponible sobre el individuo con los datos almacenados acerca de los programas educativos y ocupacionales, y utiliza todos los hechos y relaciones pertinentes para responder a las preguntas del individuo y ayudarlo a tomar una decisión. Un ejemplo de esos programas es el Sistema Interactivo de Guía e Información (*System for Interactive Guidance Information*, "SIGI", 1974–1975). El sistema, en uso por más de una década en universidades, se ha actualizado y revisado para ayudar no sólo a los estudiantes y a los adultos a prepararse para ingresar o regresar al mercado de trabajo, o para sopesar ascensos o cambios de carrera (M. R. Katz, 1993; Norris, Schott, Shatkin y Bennett, 1986).

Riesgos y directrices. Aunque es indudable que las computadoras han abierto el camino a mejoras sin precedentes en todos los aspectos de las pruebas psicológicas, algunas de sus aplicaciones pueden conducir a abusos e interpretaciones erróneas de las calificaciones de las pruebas (Butcher, 1985a; J. J. Kramer y Mitchell, 1985; Matarazzo, 1983, 1986a, 1986b). En un esfuerzo por prevenir esos riesgos se ha prestado considerable atención al desarrollo de directrices para la examinación basada en computadoras. Los *Estándares de Evaluación* (*Testing Standards*: AERA, APA, NCME, 1985) incluyen varios criterios vinculados con la evaluación basada en computadoras. Se ha elaborado un conjunto más amplio y detallado de normas que se refieren en especial al uso de la computadora en los diversos aspectos de la examinación (véase, por ejemplo, Butcher, 1987, pp. 413–431). Véase Moreland (1985, 1992), para una revisión cuidadosa del uso de computadoras en la evaluación, particularmente en relación con la interpretación computarizada de los resultados.

Dos de las preocupaciones principales respecto al uso de computadoras atañen a la comparabilidad de los resultados y a las interpretaciones narradas de los mismos. Cuando la misma prueba se aplica por medio de una computadora y de la manera impresa tradicional es necesario investigar la comparabilidad de los resultados (Mazzeo, Druesne, Raffeld, Checketts y Muhlstein, 1991). A menos que se demuestre que ambos modos de aplicación producen formas completamente igualadas de la prueba, el mismo conjunto de normas puede no ser aplicable a ambos; además, la confiabilidad y la validez del instrumento también pueden variar. Es especialmente importante revisar la comparabilidad de individuos o grupos diferentes cuya experiencia con el uso de las computadoras, sobre todo en la evaluación, puede diferir de manera notable.

El rápido crecimiento de los servicios computarizados que proporcionan informes interpretativos narrados de los resultados ha generado particular preocupación. Dos

principios básicos fundamentan muchos de los criterios. Primero, debe proporcionarse información adecuada que permita al usuario evaluar la confiabilidad, la validez y otras propiedades técnicas del sistema interpretativo empleado al preparar el programa de cómputo. ¿Cómo se obtuvieron los informes interpretativos de los resultados? ¿Cuál es la base teórica y de investigación del sistema? ¿Los análisis proceden de análisis cuantitativos o del juicio clínico de expertos? Si se trata del último caso, debe proporcionarse la información relativa a las calificaciones de estos expertos.

El segundo principio indica que siempre que se empleen informes interpretativos de los resultados (sea con propósitos clínicos o de orientación o bien cuando se utilicen al tomar decisiones importantes sobre los individuos), es indispensable considerar otras fuentes de datos disponibles sobre los examinados. Es por ello que sólo los profesionales calificados deben usar reportes de los resultados. Los reportes deben considerarse como un instrumento auxiliar y no como un sustituto del especialista profesional.

INTERPRETACIÓN DE LOS TESTS REFERIDOS A DOMINIO

Naturaleza y usos. Un acercamiento a las pruebas que generó gran actividad en los setenta, en particular en el campo de la educación, se conoció al principio como "tests referidos a criterio". Propuesto por Glaser (1963), el término se ha empleado de manera vaga y su definición varía entre diversos autores. Más aún, es común el uso de sinónimos, por ejemplo tests referidos a contenido, a un dominio o a un objetivo, aunque en ocasiones tienen connotaciones ligeramente distintas. La designación inicial de "referencia a criterio" ha sido reemplazada gradualmente por términos con mayor precisión descriptiva, por la cual en lo sucesivo emplearemos el término "tests referidos a dominio o tests de dominio".

En general, las pruebas de este tipo utilizan su marco interpretativo como referencia a un dominio específico de *contenido* más que a una población específica de *personas*. A este respecto, la aproximación ha sido comparada con la examinación usual referida a normas, que para interpretar la calificación del individuo la compara con los resultados de otros en la misma muestra; por ejemplo, en los tests referidos a un dominio la ejecución del examinado puede ser informada en términos de las operaciones aritméticas que ha dominado, el tamaño estimado de su vocabulario, el nivel de dificultad del tema de lectura que ha comprendido (desde revistas de tiras cómicas hasta la literatura clásica) o sus oportunidades de lograr un nivel determinado de ejecución sobre un criterio externo (educativo u ocupacional).

Hasta ahora, la examinación referida a dominio ha encontrado su principal aplicación en varias innovaciones educativas. Entre ellas destacan los sistemas educativos asistidos por computadoras, administrados por computadoras y otros sistemas educativos individualizados. En todos esos sistemas, la examinación está muy integrada a la instrucción y se aplica antes, durante y después de cada unidad para supervisar las habilidades que son requisitos, diagnosticar las posibles dificultades de aprendizaje y prescribir los procedimientos educativos subsecuentes (Nitko, 1989).

Desde otro ángulo, los tests referidos a dominio han sido utilizados en estudios de rendimiento educativo, como la Evaluación Nacional del Progreso Educativo (*National Assessment of Educational Progress*, E. G. Johnson, 1992; Messick, Beaton y Lord, 1983; F. B. Womer, 1970), y al hacer frente a las demandas de responsabilidad educativa. To-

davía desde otro punto de vista, otra muestra de los tests referidos a dominio se ilustra en los exámenes para el cumplimiento de las condiciones mínimas requeridas, por ejemplo para obtener una licencia de conducción o de piloto. Una aplicación relacionada es el examen de la pericia para el trabajo, que evalúa la maestría de un pequeño número de habilidades bien definidas, como en las ocupaciones militares especializadas (Maier y Hirshfeld, 1978; Swezey y Pearlstein, 1975).

Por último, la familiaridad con los conceptos de los tests referidos a dominio puede contribuir a la mejora de los exámenes tradicionales informales preparados por los maestros para su uso en el aula. Linn y Gronlund (1995) ofrecen una guía detallada para este propósito, así como un tratamiento sencillo y equilibrado de tales pruebas. Una reseña breve y excelente de las principales limitaciones de esa forma de examinación se encuentra en Ebel (1972); y Berk (1984a) proporciona una amplia revisión de muchos de los problemas técnicos de su elaboración y evaluación.

Significado del contenido. El rasgo distintivo principal de la examinación referida a dominio (como quiera que se le defina y sea que se le designe con este término o con alguno de sus sinónimos) es su interpretación del desempeño en la prueba en términos del significado del contenido. Desde luego, el punto de atención está en lo *que* los examinados pueden hacer y lo que conocen, y no en cómo se comparan con otros. Al elaborar estos instrumentos, un requisito fundamental es contar con un dominio definido con claridad de los conocimientos o las habilidades que la prueba debe evaluar. Para que los resultados tengan un significado que pueda comunicarse, el contenido del dominio por muestrearse debe ser reconocido como importante. El dominio seleccionado debe subdividirse luego en unidades pequeñas definidas en términos de ejecución. En el contexto educativo, las unidades corresponden a los objetivos educativos definidos conductualmente, como "multiplicar números de tres dígitos por otro de dos dígitos" o "identificar las faltas ortográficas que consisten en escribir s en lugar de c". En los programas preparados para la instrucción individualizada, pueden prepararse cientos de esos objetivos para una sola materia. Después de que se han formulado los objetivos educativos, se preparan los reactivos para muestrear cada uno, un procedimiento difícil y tardado; sin embargo, si no se cuenta con una especificación cuidadosa y controlada del contenido, los resultados de esta examinación pueden degenerar en una mescolanza idiosincrásica e ininterpretable. Un compromiso práctico consiste en que un experto identifique y defina los principales conceptos, principios, metodologías u objetivos educativos; cada uno de los dominios significativos definidos puede entonces muestrearse concienzudamente con los reactivos de prueba apropiados. Es indudable que el grado de especificidad con que se requiere evaluar los dominios de conducta varía según la naturaleza y propósito de la prueba (Popham, 1984; Roid, 1984).

Cuando se aplica estrictamente, la examinación referida a dominio se adapta mejor al examen de habilidades básicas (como lectura y aritmética) a nivel elemental. En esas áreas los objetivos educativos suelen disponerse en orden jerárquico, y la adquisición de las habilidades más elementales es un requisito para la adquisición de otras de mayor nivel;[7] sin embargo, no es práctico —y quizá tampoco deseable— formular objetivos muy específicos para los niveles avanzados de conocimiento en mate-

[7] Idealmente, estas pruebas siguen el modelo simplex de la escala de Guttman (véase Popham y Husek, 1969), como es el caso de las escalas ordinales de Piaget que estudiaremos en el capítulo 9.

rias menos estructuradas. A estos niveles, es probable que tanto el contenido como la secuencia de aprendizaje sean mucho más flexibles.

Por otro lado, al subrayar el significado del contenido en la interpretación de los resultados, la examinación referida a dominio puede ejercer un efecto saludable sobre la aplicación general de pruebas, por ejemplo para la interpretación de los resultados de un test de inteligencia. Describir la ejecución de un niño en un test de inteligencia en términos de las habilidades intelectuales específicas y del conocimiento que representa puede ayudar a contrarrestar las confusiones y los errores que han quedado vinculados al CI tradicional; de todas formas, cuando se plantea en esos términos generales, la aproximación referida a dominio equivale a interpretar las calificaciones de las pruebas a la luz de la validez demostrada de la prueba particular, más que en términos de vagas entidades subyacentes. Por supuesto, dicha interpretación puede combinarse con calificaciones referidas a normas.

Pruebas de destreza. Otro rasgo que suele asociarse con la evaluación referida a dominio es el procedimiento para examinar la destreza. En esencia, el procedimiento da una calificación de todo o nada que indica si el individuo ha alcanzado o no el nivel establecido de habilidad. Cuando se examinan habilidades básicas, en general se espera una destreza casi completa (por ejemplo, 80–85 por ciento de respuestas correctas). También puede emplearse una distinción de tres niveles que incluye destreza, no destreza y un intervalo intermedio dudoso o de "revisión".

En relación con la instrucción individualizada, algunos educadores han afirmado que si se cuenta con tiempo suficiente y métodos educativos apropiados, casi cualquiera puede lograr la completa destreza de los objetivos elegidos. Las diferencias individuales se manifestarían en el tiempo requerido para el aprendizaje más que en el rendimiento final, como sucede en los exámenes tradicionales (Carroll, 1963, 1970; Cooley y Glaser, 1969; Gagné, 1965). De todo esto se dice que en las pruebas de destreza las diferencias individuales son de poco o ningún interés. En consecuencia, se acostumbra elaborar los tests referidos a dominio de modo que, luego de la capacitación apropiada, minimicen las diferencias individuales en la ejecución. La evaluación de la destreza se emplea regularmente en los programas de instrucción personalizada, y también es característica de los tests referidos a dominio para habilidades básicas, adecuadas para la instrucción elemental.

En la elaboración de esas pruebas hay dos preguntas importantes: (1) ¿cuántos reactivos deben usarse para la evaluación confiable de cada objetivo educativo cubierto por la prueba?, y (2) ¿qué proporción de reactivos hay que responder correctamente para determinar en forma confiable la destreza? En buena parte de los primeros tests referidos a dominio, decisiones de criterio daban las respuestas; no obstante, el progreso sustancial alcanzado en el desarrollo de técnicas estadísticas puede proporcionar respuestas objetivas empíricas (Berk, 1984a; R. L. Ferguson y Novick, 1973; Hambleton, 1984a, 1989; Hambleton y Novick, 1973). Bastará con algunos ejemplos para ilustrar la naturaleza y el alcance de esos esfuerzos.

Estas dos preguntas pueden plantearse en una sola hipótesis que se prueba en el marco de la teoría de la toma de decisiones y el de un análisis secuencial (Hambleton, 1984a; Wald, 1947). En concreto, deseamos probar la hipótesis de que el examinado ha alcanzado el nivel requerido de destreza en el contenido del dominio u objetivo educativo muestreado por los reactivos de la prueba. El análisis secuencial consiste en hacer

observaciones, una tras otra, y decidir después de cada una si (1) se acepta la hipótesis, (2) se rechaza, o (3) hay que hacer otras observaciones. Así, el número de observaciones (en este caso, el número de reactivos) necesario para llegar a una conclusión confiable se determina durante el proceso de evaluación. En lugar de presentar un número fijo y predeterminado de reactivos, el examinado continúa resolviendo la prueba hasta que se alcance una decisión de destreza o no destreza. En este momento se interrumpe el examen y el estudiante es dirigido al siguiente nivel de instrucción o se le regresa a estudiar el nivel no dominado. Con las facilidades de las computadoras, los procedimientos de decisión secuencial son factibles y pueden reducir el tiempo total de examinación a la vez que producen estimaciones confiables de la destreza.

Algunos investigadores han explorado los métodos bayesianos de estimación que incorporan datos colaterales y son adecuados para las decisiones que requieren las pruebas de destreza. Debido al gran número de objetivos educativos que hay que probar, en general los tests referidos a dominio incluyen un pequeño número de reactivos por cada objetivo. Para completar esta limitada información se han establecido procedimientos que comprenden datos colaterales de la historia del desempeño del estudiante así como los resultados de otros alumnos (R. L. Ferguson y Novick, 1973; Hambleton, 1984a; Hambleton y Novick, 1973).

Cuando los procedimientos confeccionados individualmente no resultan prácticos, es posible establecer de manera empírica puntuaciones de corte al analizar los resultados de grupos apropiados antes y después de la instrucción. Luego se selecciona la puntuación de corte que haga una mejor discriminación entre los que han recibido la educación adecuada y los que no lo han hecho (Panell y Laabs, 1979; L. A. Shepard, 1984). En ciertas situaciones se requiere de juicio para evaluar la relativa gravedad de "aprobar" a una persona no calificada contra "reprobar" a alguien que sí lo está. Entonces, se elevaría o disminuiría el punto de corte para responder a las consecuencias de la mala clasificación.

Tests referidos a normas. Más allá de las habilidades básicas, las pruebas de destreza son inaplicables o insuficientes, ya que en las materias más avanzadas y menos estructuradas el rendimiento es abierto. El individuo puede progresar casi ilimitadamente en funciones como la comprensión, el pensamiento crítico, la apreciación y la originalidad. Más aún, la cobertura del contenido puede seguir muchas direcciones diferentes dependiendo de las habilidades, los intereses y las metas del individuo, así como de las facilidades locales de educación. En esas condiciones, la destreza total es poco realista e innecesaria, por lo que en tales casos suele utilizarse la evaluación referida a normas para determinar el grado de rendimiento. Se han elaborado algunos instrumentos para permitir tanto las aplicaciones referidas a normas como las referidas a dominio. Un ejemplo se encuentra en las pruebas Stanford de diagnóstico para lectura y matemáticas, que además de proporcionar normas apropiadas para cada nivel, permiten un análisis cuantitativo del aprovechamiento del niño en objetivos educativos detallados.

Observe que la evaluación referida a dominio no es tan nueva ni tan distinta de los exámenes referidos a normas como implican algunos de sus exponentes. Evaluar la ejecución del individuo en una prueba en términos absolutos, como una letra o el porcentaje de aciertos, desde luego es más antiguo que las interpretaciones normativas. Tentativas más precisas por describir el desempeño en una prueba en términos

del significado del contenido también son anteriores a la introducción del término "test referido a criterio" (Ebel, 1962; J. C. Flanagan, 1962; Nitko, 1984, pp. 14–16). Otros ejemplos se hallan en las primeras escalas formadas para evaluar la calidad de la escritura, la redacción o el dibujo al igualar la muestra de trabajo del individuo contra un espécimen estándar. Ebel (1972) observó además que el concepto de destreza o habilidad en educación —en el sentido del aprendizaje de todo o nada de unidades específicas— fue muy popular en los veinte y los treinta, antes de ser abandonado.

En todas las formas de evaluación está implícito un marco normativo, independientemente de cómo se expresen los resultados (Angoff, 1974; Nitko, 1984). La mera elección del contenido o bien las habilidades por medir está influida por el conocimiento del examinador de lo que puede esperarse de seres humanos en determinada etapa educativa o de desarrollo. Dicha elección presupone que se cuenta con información acerca de lo que otras personas han hecho en situaciones similares. Más aún, al imponer puntuaciones de corte uniformes sobre un continuo de habilidad, las pruebas de destreza no eliminan las diferencias individuales. Describir el nivel de lectura de comprensión de un individuo como "la habilidad de comprender el contenido del diario *The New York Times*" deja espacio para una amplia gama de diferencias individuales en grado y comprensión. Aplicar un punto de corte para establecer una dicotomía en la ejecución simplemente ignora las diferencias individuales que permanecen dentro de las dos categorías y descarta información potencialmente útil.

CALIFICACIONES MÍNIMAS Y PUNTUACIONES DE CORTE

Necesidades prácticas y peligros latentes. El concepto de destreza en la examinación referida a dominio es sólo un ejemplo del uso práctico de los puntos de corte en la toma de decisiones. En la vida cotidiana deben especificarse y adoptarse calificaciones mínimas para una gran variedad de propósitos. En muchas situaciones, las condiciones de seguridad requieren el establecimiento de puntos de corte en el desempeño, como para conceder licencias de conductor, la selección de pilotos para una aerolínea o de trabajadores para una planta nuclear. En la escuela, aprobar un curso o graduarse representan otras situaciones que exigen una clasificación de todo o nada (Jaeger, 1989). En la práctica clínica y de consejería, las decisiones concernientes al tratamiento o las recomendaciones pueden requerir de juicios similares.

Un argumento particularmente sólido para el uso de los puntos de corte atañe a la presencia de variables críticas requeridas para el desempeño de ciertas funciones, se trata de variables en las que una deficiencia no puede compensarla una habilidad sobresaliente en otras variables. En tales casos, una calificación alta en una batería de selección podría ocultar una deficiencia en una habilidad crucial. Con todo, los individuos que caen debajo del mínimo requerido en la habilidad esencial fallarán, independientemente de sus otras habilidades; por ejemplo, los operadores del equipo de sonar deben tener una buena discriminación auditiva. Durante la Segunda Guerra Mundial, los reclutas de la marina estadounidense fueron seleccionados para el entrenamiento como operadores de sonar sobre la base de la combinación de sus calificaciones en pruebas de discriminación auditiva y de comprensión mecánica. Como resultado, algunos hombres con formación universitaria que obtuvieron excelentes resultados en comprensión mecánica pero que resultaron deficientes en las habilidades auditivas esenciales fueron asignados a dicho entrenamiento con el fracaso subsecuente. El procedimiento estándar de la marina requería que quienes fracasaran en su

primera asignación de entrenamiento fueran transferidos a tareas generales como aprendices de marinero, lo que suponía la pérdida de sus servicios especializados. Un análisis posterior de la situación llevó a la sustitución de un procedimiento de corte para este propósito de selección; sin embargo, para la mayor parte de las variables relacionadas con el trabajo, la relación con la ejecución del trabajo tiende a ser lineal, de modo que entre mayor sea el resultado obtenido en la prueba mejor suele ser el desempeño (Coward y Sackett, 1990). En tales casos, la calificación real del sujeto predice mejor su rendimiento que su posición relativa en un punto de corte.

En la medida en que no es posible evitar el uso de los puntos de corte en muchas decisiones prácticas, es esencial estar al tanto de los riesgos de dichas evaluaciones y utilizar procedimientos para reducir los errores de juicio; por ejemplo, deben hacerse esfuerzos para mitigar las limitaciones del resultado de una sola prueba. Cuando sea factible, el corte debe ser una banda de puntuaciones más que una sola puntuación obtenida en una aplicación de la prueba. Más aún, las decisiones sobre los individuos deben depender de múltiples fuentes de información, de modo que los resultados de la prueba los completen otros datos relevantes a su desempeño, pasado y actual. Si un grupo de jueces establece los puntos de corte, éste debe incluir la representación adecuada de expertos tanto en el área relativa al desempeño de la tarea como en los principios de construcción y uso de las pruebas. Por encima de todo, siempre que sea posible, los puntos de corte deben establecerse o verificarse sobre la base de datos empíricos. En concreto, esto implica que deben obtenerse los resultados conseguidos en la prueba por grupos que difieren claramente en la conducta de criterio importante, como el desempeño real en cierto trabajo. Por supuesto, esta es la ejecución que la prueba pretende predecir y en que se establece un punto de corte para asegurar un mínimo seguro, aceptable o deseable. Un ejemplo claro de la aplicación del método empírico para establecer puntuaciones de corte en una prueba de selección de personal se encuentra en las tablas de expectabilidad que veremos en la siguiente sección.

Tablas de expectabilidad. Una de las formas de interpretar el significado del resultado de una prueba es la ejecución criterio esperada del sujeto, como en el programa de capacitación o en el empleo. Este uso del término "criterio" sigue la práctica psicométrica habitual, como cuando se dice que una prueba es validada contra un criterio particular (véase el capítulo 1). Hablando estrictamente, el termino "test referido al criterio" debe aludir a esta interpretación del desempeño, mientras que los otros planteamientos que revisamos en la sección anterior se describen mejor como referidos a contenido o a dominio.

La tabla de expectabilidad muestra la probabilidad de diferentes resultados de criterio para quienes obtienen cada resultado de la prueba; por ejemplo, si un estudiante obtuvo una calificación de 530 en la Prueba de Evaluación Académica (*Scholastic Assessment Test*, SAT) de la Junta Universitaria, ¿cuáles son las posibilidades de que su calificación promedio en determinada universidad caiga en la categoría A, B, C, D o F? Esta información se puede obtener si se examina la distribución bivariada de las calificaciones predictoras (del SAT) contra el criterio establecido (el promedio de calificaciones). Si el número de casos en cada celda de dicha distribución bivariada se cambia por un porcentaje, el resultado es una tabla de expectabilidad, como la que ilustra la tabla 3.6. Los datos para esta tabla fueron obtenidos de 211 estudiantes de séptimo grado inscritos en cursos de matemáticas. El predictor fue la prueba de Razonamiento Numérico de los Tests de Aptitud Diferencial (*Differential Aptitude Tests*, DAT) administrados

Tabla 3.6

Tabla de expectabilidad que muestra la relación entre la Prueba de razonamiento numérico (DAT) y las calificaciones en un curso de matemáticas para 211 estudiantes de séptimo grado

Puntuación de la prueba	Número de casos	Porcentaje que recibe cada calificación o grado por letra			
		D y menos	C	B	A
30 y por encima	22	5	0	36	59
20–29	104	9	21	43	27
10–19	71	37	37	24	3
Menos de 10	14	43	36	14	7

(Adaptado de *Technical Manual for Differential Aptitude Tests*, 5ª ed., p. 152. Reproducido con autorización. Copyright " 1992 por la Psychological Corporation.)

al final del primer semestre. El criterio fue las calificaciones finales del segundo semestre, la correlación entre los resultados de la prueba y el criterio fue de .60.

La primera columna de la tabla 3.6 muestra las calificaciones de la prueba divididas en cuatro intervalos de clase; en la segunda columna se presenta el número de estudiantes cuyas puntuaciones caen en cada intervalo. Las entradas restantes en cada hilera de la tabla indican el porcentaje de casos dentro de cada intervalo de calificaciones que recibió cada grado al final del curso. Así, de los 22 estudiantes con calificaciones iguales o superiores a 30 en la prueba de razonamiento numérico, cinco por ciento recibió calificaciones iguales o inferiores a D, ninguno recibió C, 36 por ciento recibió B y 59 por ciento A. En el otro extremo, de los 14 estudiantes con una calificación inferior a 10 en la prueba, 43 por ciento recibió calificaciones iguales o inferiores a D, 36 por ciento C y 14 por ciento B. El anómalo siete por ciento que recibió A, que representa un solo caso, finalmente es información inútil para propósitos de generalización, como es el caso del cinco por ciento con calificaciones iguales o superiores a 30 que recibió una calificación igual o menor a D, nuevamente representado por un caso. No obstante, dentro de las limitaciones de los datos disponibles, los porcentajes de la tabla 3.6 proporcionan estimaciones de la probabilidad de que un individuo recibirá una determinada calificación criterio. Por ejemplo, si un nuevo estudiante recibe una puntuación de 24 en la prueba de razonamiento numérico del DAT (es decir, cae en el intervalo 20–29), debemos concluir que la probabilidad de que obtenga en el curso una calificación de A es de 27 por ciento, la de que obtenga una calificación de B es 43 por ciento, etcétera.

En muchas situaciones prácticas, los criterios pueden dicotomizarse como "éxito" y "fracaso" en un trabajo, un curso o en otra empresa. En tales condiciones, puede prepararse una *gráfica de expectabilidad* que muestre la probabilidad de éxito o fracaso que corresponde a cada intervalo de calificación. La figura 3.7 es un ejemplo de dicha gráfica, la cual se basa en una batería para la selección de pilotos elaborada por la fuerza aérea de los Estados Unidos. La gráfica muestra el porcentaje de cadetes que califican dentro de cada estanina de la batería y que no lograron completar el entrenamiento de

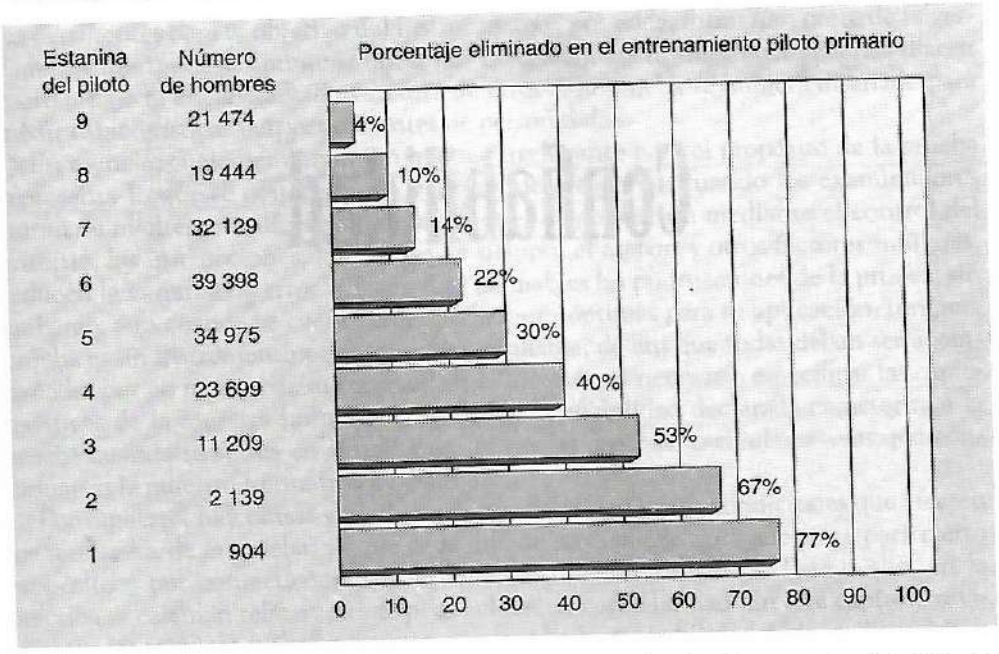


Figura 3.7. Gráfica de Expectabilidad que muestra la relación entre la ejecución en una batería para la selección de pilotos y la eliminación de un vuelo de entrenamiento primario.

(Tomado de J. C. Flanagan, 1947, p. 58.)

vuelo. Puede verse que el 77 por ciento de los cadetes que recibieron una estanina de 1 fue eliminado del curso de entrenamiento, mientras que sólo cuatro por ciento de los que obtuvieron una estanina 9 no logró completar el entrenamiento satisfactoriamente. Entre esos extremos, el porcentaje de fracasos disminuyó en forma constante a lo largo de las estaninas sucesivas. Sobre la base de esta gráfica de expectabilidades puede predecirse, por ejemplo, que alrededor del 40 por ciento de los cadetes que obtienen una calificación estanina de 4 fracasará y que cerca del 60 por ciento completará satisfactoriamente el entrenamiento de vuelo. Pueden hacerse afirmaciones similares acerca de la probabilidad de éxito o fracaso de los individuos que reciben cada estanina. Así, la probabilidad de que complete el entrenamiento un individuo con una calificación estanina de 4 es de 60:40 o 3:2. Además de brindar una interpretación referida al criterio de los resultados de la prueba, puede observarse que tanto las tablas como las gráficas de expectabilidad dan una idea general de la validez de una prueba para predecir un criterio determinado. Es por ello que al final del capítulo 6, en una sección sobre los modelos de decisión para el uso justo de las pruebas, haremos una revisión más detallada de los procedimientos empíricos empleados para establecer puntuaciones de corte. En dicha sección también nos referimos a los procedimientos matemáticos para el establecimiento de calificaciones de corte óptimas en diversas condiciones. Las aplicaciones específicas de las puntuaciones de corte en las principales áreas de la práctica psicológica también pueden encontrarse en el capítulo 17.

Confiabilidad

El término "confiabilidad" se refiere a la consistencia de las puntuaciones obtenidas por las mismas personas cuando se las examina en distintas ocasiones con el mismo test, con conjuntos equivalentes de reactivos o en otras condiciones de examinación. El concepto fundamenta el cálculo del *error de medición* de un solo resultado, con el que podemos predecir la probable fluctuación en la calificación de un solo individuo debida a factores aleatorios irrelevantes o desconocidos.

El concepto de confiabilidad se ha dirigido a varios aspectos de la consistencia de las calificaciones. En su sentido más amplio, la confiabilidad de una prueba indica la medida en que las diferencias individuales en los resultados pueden atribuirse a "verdaderas diferencias" en las características consideradas y el grado en que pueden deberse a errores fortuitos. En términos más técnicos, la medición de la confiabilidad de una prueba permite estimar qué proporción de la varianza total de las puntuaciones se debe a la *varianza de error*. No se trata de "errores" en el sentido de que hubieran podido evitarse o corregirse con una mejor metodología. El sustantivo *error* es una herencia de las primeras épocas de la psicología, cuando estaba interesada en encontrar las leyes generales de la conducta y en evaluar lo que se suponía que eran rasgos básicos fijos. En la actualidad, los psicólogos reconocen la variabilidad como una propiedad intrínseca del comportamiento y procuran investigar e identificar sus muchas fuentes.

Con respecto a la confiabilidad de las puntuaciones, el meollo del asunto estriba en la definición de la varianza de error. Factores que para un propósito podrían considerarse varianza de error para otro serían clasificados bajo el rubro de varianza verdadera; por ejemplo, si nos interesa medir las fluctuaciones del estado de ánimo, los cambios registrados diariamente en las puntuaciones de una prueba de alegría-depresión serían

los pertinentes para el objetivo del instrumento y, por ende, formarían parte de la verdadera varianza de las puntuaciones. Por otro lado, las mismas fluctuaciones diarias caerían bajo el encabezado de varianza de error si la prueba se hubiera diseñado para medir características más permanentes de personalidad.

En esencia, cualquier condición que sea irrelevante para el propósito de la prueba representa lo que se conoce como varianza de error. Así, cuando los examinadores tratan de mantener uniformes las condiciones de aplicación mediante el control del entorno, las instrucciones, los límites de tiempo, el *rapport* y otros factores similares, reducen la varianza de error y hacen más confiables las puntuaciones de la prueba; sin embargo, aun cuando se cuente con condiciones óptimas para su aplicación, ninguna prueba es un instrumento perfectamente confiable, de ahí que todas deban ser acompañadas por un reporte de su confiabilidad. También es necesario especificar las características de la muestra normativa, ya que la confiabilidad declarada caracteriza a la prueba cuando se aplica en condiciones normales a personas similares a las que conformaron la muestra normativa u original.

Por supuesto, hay tantas variedades de confiabilidad como condiciones que afecten los resultados de la prueba, ya que cada una de éstas puede ser irrelevante para cierto propósito y por consecuencia clasificada como varianza de error. Pese a ello, en la práctica se calculan relativamente pocas clases de confiabilidad. En este capítulo revisaremos las técnicas más importantes para medir la confiabilidad de las puntuaciones junto con las fuentes de varianza de error identificadas para cada una.[1]

Todas las categorías de confiabilidad pueden expresarse en términos de un *coeficiente de correlación* porque a todas les atañe el grado de consistencia o acuerdo entre dos conjuntos de puntuaciones obtenidas independientemente; en consecuencia, en la siguiente sección consideramos algunas de las características básicas de los coeficientes de correlación para aclarar su uso e interpretación. En cualquier texto básico de estadística para la educación o la psicología, como los trabajos de Runyon y Haber (1991) o de D. C. Howell (1997), es posible encontrar una revisión más técnica de la correlación y una especificación más detallada de los procedimientos para calcularla.

EL COEFICIENTE DE CORRELACIÓN

Significado de la correlación. En esencia, un coeficiente de correlación (r) expresa el grado de correspondencia, o *relación*, entre dos conjuntos de puntuaciones. Si el individuo que obtiene la mayor calificación en la variable 1 consigue el mejor resultado en la variable 2, el individuo con la segunda mejor calificación en la variable 1 también logra el segundo mejor resultado en la variable 2, y así va disminuyendo hasta llegar al individuo con la menor calificación del grupo, entonces existiría una correlación perfecta entre las variables 1 y 2 y sería igual a +1.00.

[1] Esta aproximación a la confiabilidad de las puntuaciones también se conoce con el nombre de teoría de la generalización de la confiabilidad (véase Brennan, 1994; Crick y Brennan, 1982; Cronbach, Gleser, Nanda y Rajaratnam, 1972; Feldt y Brennan, 1989; Shavelson y Webb, 1991). Sin embargo, no es una designación lo suficientemente específica, porque la generalización se aplica a todos los aspectos de la puntuación de una prueba y, de hecho, a todo dato científico. Una descripción más precisa de este procedimiento de confiabilidad se basa en su identificación de los componentes de la varianza como relevantes o irrelevantes.

En la figura 4.1 se ilustra una correlación positiva perfecta hipotética, dispuesta en un diagrama de dispersión o distribución bivariada. Cada marca en el diagrama indica la calificación de una persona tanto en la variable 1 (eje horizontal) como en la variable 2 (eje vertical). Advierta que los 100 casos del grupo se distribuyen por la diagonal que corre del extremo inferior izquierdo al extremo superior derecho. Una distribución de esa forma indica una correlación positiva perfecta (+1.00), ya que muestra que cada persona ocupa la misma posición relativa en ambas variables. Entre más se aproxime la distribución bivariada de puntuaciones a esta diagonal, más alta será la correlación positiva.

La figura 4.2 ilustra una correlación negativa perfecta (–1.00), en la que hay una completa inversión de puntuaciones de una variable a otra: el mejor individuo en la variable 1 obtuvo el peor resultado en la variable 2 y viceversa, y esta inversión se mantiene a lo largo de la distribución. Observe que todas las personas caen en la diagonal que se extiende del extremo superior izquierdo al extremo inferior derecho, es decir, que corre en dirección opuesta a la de la figura 4.1.

Una correlación de cero indica la ausencia absoluta de relación, como podría ocurrir fortuitamente. Si el nombre de cada persona fuera sacado al azar de una caja para determinar su posición en la variable 1, y se repitiera el proceso para la variable 2, se

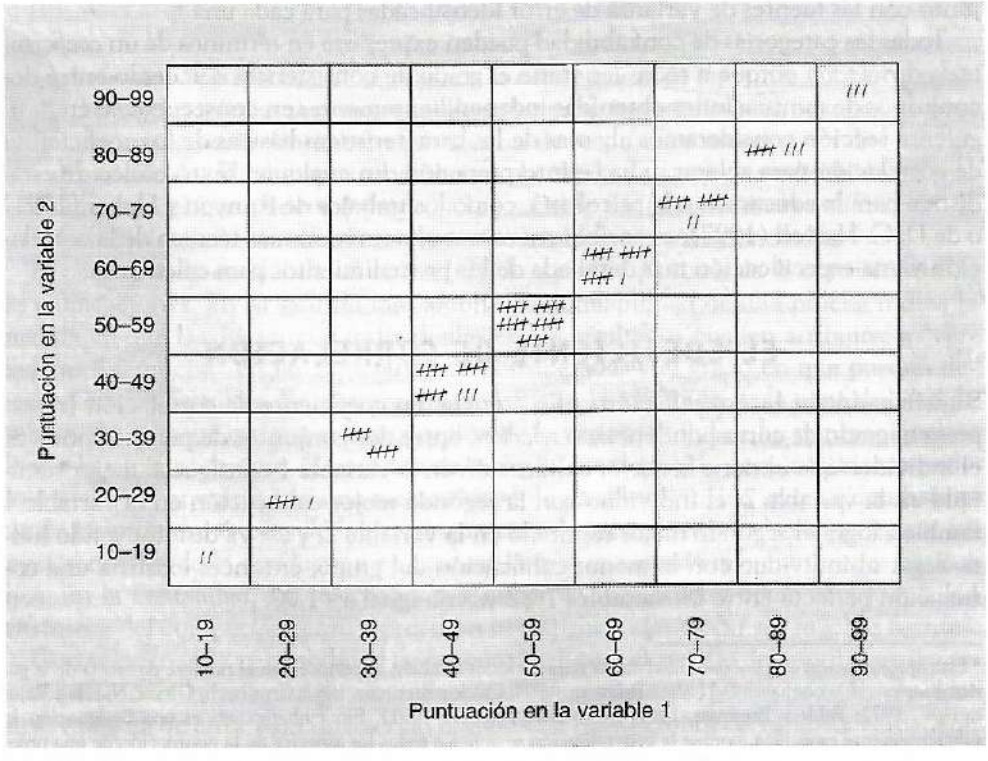


Figura 4.1. Distribución bivariada de una correlación hipotética de +1.00.

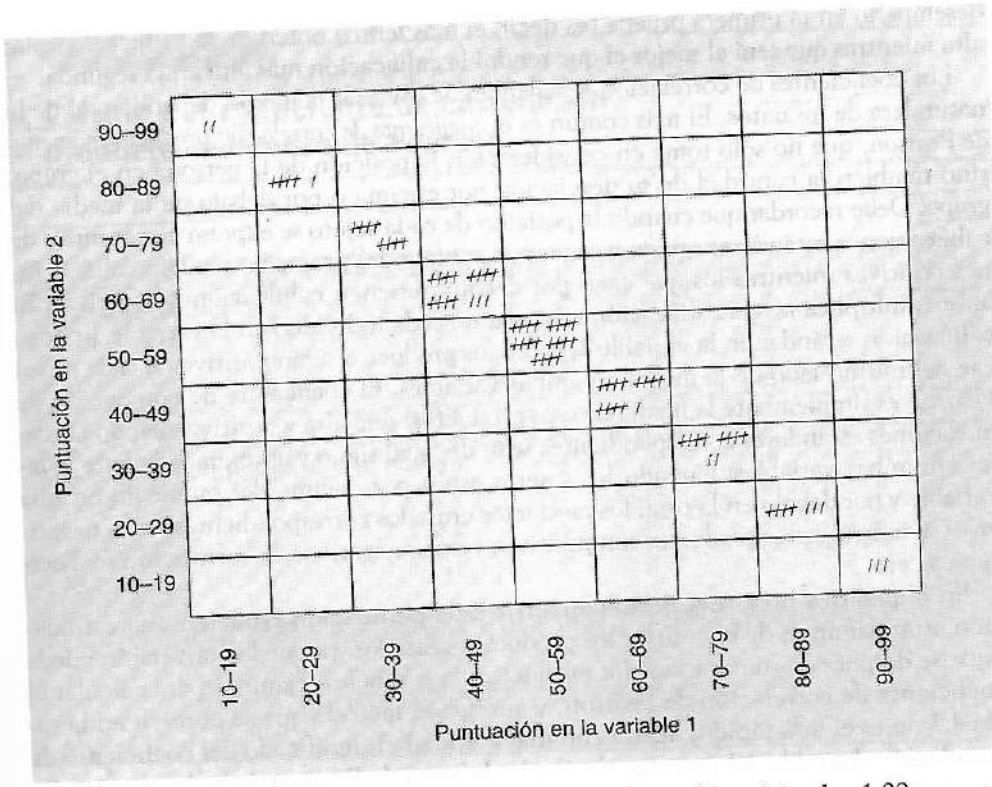


Figura 4.2. Distribución bivariada de una correlación hipotética de –1.00

encontraría una correlación igual o cercana a cero. En tales condiciones, sería imposible predecir la posición relativa de un individuo en la variable 2 conociendo su calificación en la variable 1. La persona con el mejor puntaje en la variable 1 podría obtener una puntuación alta, baja o promedio en la variable 2. Por casualidad, algunos tendrían altas puntuaciones en ambas variables, o bien inferiores al promedio; otros quedarían por arriba del promedio en una variable y por debajo en otra, mientras que otros más se ubicarían por encima del promedio en una y en el promedio en la otra, etc. No habría regularidad en la relación de un individuo a otro.

En general, los coeficientes que se encuentran en la práctica se localizan entre estos extremos y dan valores superiores a cero pero menores a 1.00. Las correlaciones entre las medidas de habilidades son casi siempre positivas, aunque a menudo son bajas. Cuando se obtiene una correlación negativa entre dos de las variables, por lo general se debe a la forma de expresar los resultados; por ejemplo, si se correlacionan calificaciones de tiempo con calificaciones de cantidad es probable que se obtenga una correlación negativa. Así, puede esperarse una correlación negativa entre los resultados de una prueba de cálculo y otra de razonamiento aritmético si la calificación de cada persona en la primera prueba se registra como el número de minutos requeridos para completar todos los reactivos mientras que la calificación de la segunda representa el número de problemas resueltos correctamente. En ese caso, el individuo con el peor

desempeño en la primera prueba (es decir, el más lento) obtendrá la calificación más alta mientras que será el mejor el que tendrá la calificación más alta en la segunda.

Los coeficientes de correlación se calculan de diversas maneras, dependiendo de la naturaleza de los datos. El más común es el *coeficiente de correlación producto-momento de Pearson*, que no sólo toma en consideración la posición de la persona en el grupo, sino también la cantidad de su desviación por encima o por debajo de la media del grupo. Debe recordar que cuando la posición de cada sujeto se expresa en términos de calificaciones estándares, quienes caen por encima del promedio reciben calificaciones positivas mientras los que caen por debajo obtienen calificaciones negativas. Si ahora multiplicamos la calificación estándar de cada individuo en la variable 1 por su calificación estándar en la variable 2, todos los productos serán positivos si cada quien cae del mismo lado de la media en ambas variables. El coeficiente de correlación de Pearson es simplemente la media de esos productos: será alta y positiva cuando las calificaciones estándares correspondientes sean de igual signo y aproximadamente iguales en ambas variables. Cuando los sujetos estén por encima del promedio en una variable y por debajo en la otra, los productos cruzados correspondientes serán negativos. Cuando algunos productos son positivos y otros negativos, la correlación será cercana a cero.

En la práctica no es necesario convertir cada puntuación cruda en una calificación estándar antes de encontrar los productos cruzados, ya que la conversión puede hacerse después de sumar todos los productos. Hay muchos caminos para calcular el coeficiente de correlación de Pearson, y aunque el método que se observa en la tabla 4.1 no es el más rápido, ilustra con más claridad el significado del coeficiente de correlación. La tabla 4.1 muestra el cálculo de una r de Pearson entre las puntuaciones de 10 niños en matemáticas (X) y en lectura (Y). Junto al nombre de cada uno se encuentran sus calificaciones. Las sumas y las medias de las 10 puntuaciones se presentan bajo las columnas respectivas; la tercera columna muestra la desviación (x) de cada puntuación de matemáticas respecto a su media y en la cuarta se encuentra la desviación (y) de cada puntuación de lectura de la media correspondiente. En las dos siguientes columnas cada una de las desviaciones se eleva al cuadrado, y las sumas de los cuadrados se utilizan para calcular las desviaciones estándares de las calificaciones de matemáticas y de lectura con el método descrito en el capítulo 3. Para encontrar las calificaciones estándares no dividimos cada x y y entre su DE correspondiente, sino que realizamos la división una sola vez al final, como lo indica la fórmula de correlación de la tabla 4.1. Los productos cruzados de la última columna (xy) se obtienen multiplicando las desviaciones correspondientes en las columnas x y y. Para calcular la correlación (r), se divide la suma de estos productos cruzados entre el producto del número de casos (N) por las dos desviaciones estándares ($DE_x DE_y$).

Significancia estadística. En la tabla 4.1 se encontró una correlación de .40 que indica una relación positiva moderada entre las puntuaciones de matemáticas y lectura. Hay cierta tendencia a que los niños que tuvieron buen desempeño en matemáticas también lo tuvieran en la prueba de lectura y viceversa, aunque la relación no es estrecha. Si sólo estuviésemos interesados en la ejecución de los 10 niños podría-

Tabla 4.1

Cálculo del coeficiente de correlación producto-momento de Pearson

Alumno	Matemáticas X	Lectura Y	x	y	x^2	y^2	xy
Guillermo	41	17	+1	−4	1	16	− 4
Carolina	38	28	−2	+7	4	49	−14
Alfredo	48	22	+8	+1	64	1	8
Ana	32	16	−8	−5	64	25	40
Roberto	34	18	−6	−3	36	9	18
Juana	36	15	−4	−6	16	36	24
Elena	41	24	+1	+3	1	9	3
Ruth	43	20	+3	−1	9	1	− 3
Ricardo	47	23	+7	+2	49	4	14
María	40	27	0	+6	0	36	0
Σ	400	210	0	0	244	186	86
M	40	21					

$$DE_x = \sqrt{\frac{244}{10}} = \sqrt{24.40} = 4.94 \qquad DE_y = \sqrt{\frac{186}{10}} = \sqrt{18.60} = 4.31$$

$$r_{xy} = \frac{\Sigma xy}{(N)(DE_x)(DE_y)} = \frac{86}{(10)(4.94)(4.31)} = \frac{86}{212.91} = .40$$

mos aceptar esta correlación como una descripción adecuada del grado de relación que existe entre las dos variables en el grupo, pero por lo general la investigación psicológica pretende generalizar más allá de la *muestra* particular de individuos examinados a la *población* que representan; por ejemplo, quizá desee saber si las habilidades para las matemáticas y la lectura están correlacionadas entre los escolares de la misma edad de los niños a los que se aplicaron las pruebas, en cuyo caso es obvio que los 10 sujetos examinados constituyen una muestra inadecuada de esta población. Otra muestra comparable del mismo tamaño arrojaría una correlación mucho mayor o mucho menor.

Los investigadores cuentan con procedimientos estadísticos para estimar la probable fluctuación de una muestra a otra en el tamaño de las correlaciones, medias, desviaciones estándares y cualquier otra medida. Pero la pregunta que generalmente se plantea es si las correlaciones son significativamente mayores a cero. En otras palabras, si en la población la correlación es de cero, ¿una correlación tan alta como la de nuestra muestra podría ser el resultado sólo de un error de muestreo? Cuando se dice que una correlación es "significativa al nivel del uno por ciento (.01)" significa que no hay más de una oportunidad en 100 de que la correlación de la población sea de cero, por lo que se concluye que las dos variables están verdaderamente correlacionadas. Los niveles de significancia se refieren al riesgo de error que estamos dispuestos a correr al sacar conclusiones de nuestros datos. Si se dice que una correlación es significativa al nivel de .05, la probabilidad de error es de cinco en cada 100. La mayor parte de la

investigación psicológica utiliza los niveles de .01 o de .05, aunque pueden emplearse otros por razones especiales.

Como era de esperarse, la correlación de .40 de la tabla 4.1 no llega a ser significativa al nivel de .05, pues con sólo 10 casos es difícil establecer de manera concluyente una relación general. Con una muestra de este tamaño, la menor correlación significativa a nivel de .05 es .63. Cualquier correlación menor no permite responder la pregunta de si las dos variables están correlacionadas en la población de la que se extrajo la muestra. Las correlaciones significativas mínimas a los niveles de .01 y .05 para grupos de diferentes tamaños pueden encontrarse en cualquier libro de estadística al consultar las tablas de significancia de las correlaciones; sin embargo, para los propósitos de interpretación de este libro lo único que se requiere es una comprensión general del concepto.

Durante muchos años, los niveles de significancia han sido la forma tradicional de evaluar las correlaciones, pero cada vez es mayor la conciencia de la insuficiencia y los defectos de este procedimiento. Demostrar que un coeficiente de confiabilidad (o cualquier correlación) es significativamente mayor a cero proporciona pocos conocimientos para propósitos teóricos o prácticos. Ni siquiera una correlación elevada logra pasar la "prueba de significancia" cuando se obtiene de una muestra tan pequeña. Una posible alternativa consiste en considerar la magnitud real de la correlación obtenida y estimar el *intervalo de confianza* dentro del cual es probable que caiga la correlación de la población a un nivel especificado de confianza (véase, por ejemplo, Carver, 1993; J. Cohen, 1994; Hunter y Schmidt, 1990; Olkin y Finn, 1995; Schmidt, 1996; W. W. Tryon, 1996). Esta tendencia al uso de los intervalos de confianza, como complemento si no es que como sustituto de las pruebas de significancia, anuncia un cambio importante en el análisis de los coeficientes de correlación en los años por venir.

El coeficiente de confiabilidad. Los coeficientes de correlación tienen muchos usos en el análisis de los datos psicométricos, entre ellos la medición de la confiabilidad de una prueba. En la figura 4.3 se presenta un ejemplo de un coeficiente de confiabilidad, calculado por el método producto-momento de Pearson, en el que se correlacionaron las puntuaciones obtenidas por 104 personas en dos formas equivalentes de una prueba de fluidez verbal.[2] En una forma, los examinados disponían de cinco minutos para escribir tantas palabras como pudieran que empezaran con determinada letra; la segunda forma era idéntica a la primera, excepto porque empleaba una letra diferente. Los autores del instrumento eligieron las letras de modo que su nivel de dificultad fuera aproximadamente igual para este objetivo.

Se descubrió que la correlación entre el número de palabras escritas en las dos formas de la prueba era igual a .72, una correlación elevada y significativa al nivel de .01. Con 104 casos, cualquier correlación igual o mayor a .25 es significativa a este nivel. No obstante, la correlación obtenida es algo menor a lo que se considera deseable para los coeficientes de confiabilidad, que por lo general caen en .80 o .90. Un examen del

[2] Uno de los subtests del SRA del Test de Habilidades Mentales Primarias para edades de 11 a 17. Los datos fueron obtenidos de una investigación de Anastasi y Drake (1954).

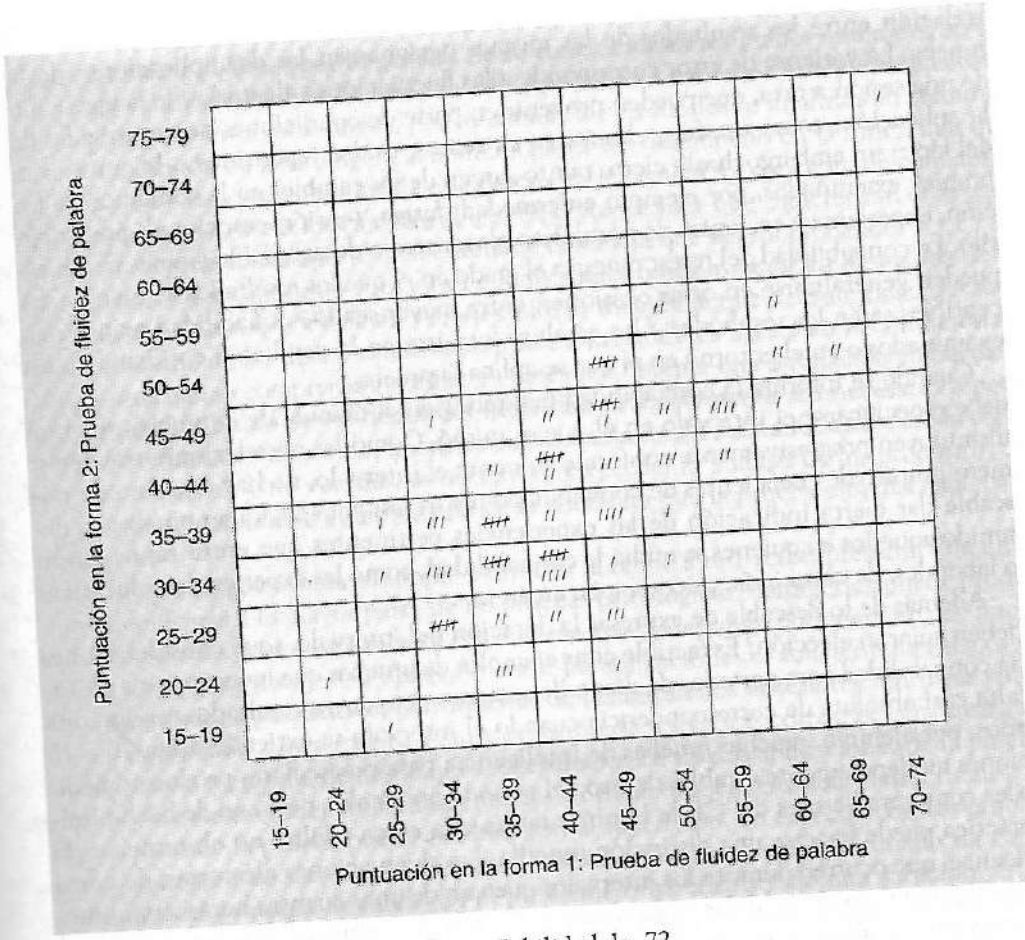


Figura 4.3. Un coeficiente de confiabilidad de .72.
(Datos de Anastasi y Drake, 1954).

diagrama de dispersión de la figura 4.3 muestra la clásica distribución bivariada de los resultados correspondientes a una correlación positiva elevada. Observe que las marcas se agrupan cerca de la diagonal que se extiende del extremo inferior izquierdo al extremo superior derecho; aunque la tendencia sigue definitivamente esta dirección, se observa cierta dispersión de entradas individuales. En la siguiente sección se considera el uso del coeficiente de correlación para calcular diferentes medidas de la confiabilidad de la prueba.

TIPOS DE CONFIABILIDAD

Confiabilidad test-retest. El método más obvio para encontrar la confiabilidad de las puntuaciones de una prueba consiste en aplicar el mismo instrumento por segunda ocasión. En este caso, el coeficiente de confiabilidad (r_{tt}) es simplemente la co-

rrelación entre los resultados de las mismas personas en las dos aplicaciones de la prueba. La varianza de error corresponde a las fluctuaciones aleatorias de la ejecución de una sesión a otra, que pueden provenir en parte de condiciones no controladas de la aplicación, como cambios climáticos extremos, ruidos repentinos o la punta rota del lápiz; sin embargo, hasta cierto punto surgen de los cambios en la condición de los propios examinados, por ejemplo enfermedad, fatiga, tensión emocional, preocupación, experiencias recientes de naturaleza placentera o desagradable y cosas por el estilo. La confiabilidad del retest muestra el grado en el que los resultados de una prueba pueden generalizarse en otras ocasiones; entre mayor sea la confiabilidad menos susceptibles serán los resultados a los cambios fortuitos en la condición cotidiana de los examinados o en el entorno en el que se aplica la prueba.

Cuando se informa la confiabilidad test-retest en el manual de la prueba, siempre debe especificarse el intervalo en el que se midió. Como las correlaciones test-retest disminuyen progresivamente conforme aumenta el intervalo, no hay uno sino un número infinito de coeficientes de confiabilidad test-retest para cualquier prueba. Es deseable dar cierta indicación de las experiencias pertinentes que entre tanto hayan tenido aquellos en quienes se midió la confiabilidad, como las experiencias educativas o laborales, de consejería, psicoterapia, etcétera.

Además de lo deseable de expresar la duración del intervalo, ¿qué consideraciones deben guiar su elección? Es factible citar ejemplos de pruebas que muestran una elevada confiabilidad para periodos de días o de semanas pero cuyos resultados revelan una falta casi absoluta de correspondencia cuando el intervalo se extiende hasta 10 o 15 años; por ejemplo, muchas pruebas de inteligencia para preescolares producen mediciones moderadamente estables dentro del periodo preescolar, pero son de hecho inútiles como predictores del CI de la niñez tardía o la etapa adulta; no obstante, en la práctica puede hacerse una distinción sencilla. Las fluctuaciones aleatorias, de corto alcance, que ocurren durante los intervalos que van de unas cuantas horas a unos meses en general están incluidas en la varianza de error del resultado de la prueba. Así, al supervisar esta clase de confiabilidad hay que hacer el esfuerzo por mantener un intervalo corto. Al examinar a niños pequeños el periodo debe ser aún más corto, ya que en las edades tempranas es posible percibir cambios progresivos del desarrollo en un lapso de un mes o incluso menos. Para cualquier persona, el intervalo entre ambas pruebas no debería exceder a los seis meses.

En cuanto a los cambios ocurridos en periodos mayores, es más probable que sean acumulativos y progresivos a que se deban por entero al azar. Más aún, es factible que caractericen un área de conducta más amplia que la que cubre el propio desempeño en la prueba. Así, el nivel general de aptitud académica, comprensión mecánica o juicio artístico puede haberse modificado apreciablemente al cabo de 10 años debido a experiencias inusuales. La posición del individuo en relación con otras personas de su edad puede haber aumentado o disminuido en forma notable debido a circunstancias peculiares del hogar, la escuela o del entorno, o a otras razones como enfermedades o perturbación emocional.

El grado en que esos factores pueden afectar el desarrollo psicológico es un importante problema que hay que investigar; pero esta cuestión no debe confundirse con el problema de la confiabilidad de determinado instrumento. Así, al medir la confiabili-

dad del Stanford-Binet por lo general no se correlacionan resultados obtenidos en un lapso de 10 años, o incluso de uno, sino de unas cuantas semanas. Aunque en efecto se han empleado intervalos largos, por lo común los resultados se analizan en términos de la predicción de la inteligencia adulta a partir del desempeño en la niñez, más que en términos de la confiabilidad de una prueba en particular. El concepto de confiabilidad suele restringirse a los cambios al azar de corto alcance que caracterizan el desempeño en la prueba más que al área entera de conducta que se está probando.

Advierta que diferentes funciones conductuales pueden variar en la medida de la fluctuación diaria que exhiban; por ejemplo, es indudable que la estabilidad del movimiento fino de los dedos es más susceptible a los cambios ligeros en la condición de la persona que su comprensión verbal. Si deseamos obtener una estimación global de la estabilidad habitual de los dedos del individuo probablemente sea necesario repetir la prueba durante varios días, mientras que una sola sesión sería suficiente para la comprensión verbal. Una vez más, es necesario retroceder al análisis de los propósitos de la prueba y a una cabal comprensión de la conducta que la prueba está diseñada para predecir.

Aunque en apariencia es simple y directa, la técnica test-retest presenta dificultades al aplicarse a la mayor parte de las pruebas psicológicas. La práctica puede producir diferentes grados de mejora en las puntuaciones que obtienen en el retest individuos distintos, y si el intervalo entre las aplicaciones es suficientemente corto, los examinados pueden recordar muchas de las respuestas anteriores. En otras palabras, es probable que se repita el mismo patrón de respuestas correctas y erróneas por efecto de la memoria, de modo que los resultados en las dos aplicaciones de la prueba no se obtuvieron independientemente y su correlación será elevada y falsa. Incluso la naturaleza de la prueba puede cambiar con la repetición, en especial con los problemas que suponen razonamiento o ingenuidad. Una vez que el examinado ha comprendido el principio incluido en el problema o ha encontrado una solución, en el futuro podrá reproducir la respuesta correcta sin los pasos intermedios. Sólo las pruebas que no son afectadas de manera apreciable por la repetición son adecuadas para la técnica test-retest. En esta categoría se encue ıtra una serie de pruebas motoras y de discriminación sensorial; sin embargo, en la mayor parte de los casos no es apropiado repetir el examen con el mismo instrur ento para encontrar el coeficiente de confiabilidad.

Confiabilidad de formas alternas. El uso de formas alternas de la prueba es una manera de evitar las dificultades de la confiabilidad test-retest. Las mismas personas pueden ser evaluadas con una forma en la primera ocasión y con otra equivalente en la segunda. La correlación entre las puntuaciones de las dos formas representa el coeficiente de confiabilidad de la prueba, que no sólo mide la estabilidad temporal, sino también la consistencia de las respuestas a diferentes muestras de reactivos (o formas de la prueba), lo que permite combinar dos tipos de confiabilidad. Como ambas son importantes para casi todos los propósitos de la examinación, la confiabilidad de formas alternas proporciona una medida útil para evaluar muchas pruebas.

Es conveniente examinar de cerca el concepto de muestreo de reactivos, o *muestreo de contenido*, ya que no sólo fundamenta la confiabilidad de formas alternas, sino tam-

bién otras clases de confiabilidad. La mayoría de los estudiantes ha vivido la experiencia de presentar un examen en el que cree haber tenido "un golpe de suerte" porque muchos de los reactivos incluían los temas que estudió con más cuidado. Otras veces pueden haber tenido la experiencia opuesta y encontrar un número inusualmente grande de reactivos sobre temas que no revisaron. Esta conocida situación ilustra la varianza de error que resulta del muestreo de contenido. ¿En qué medida dependen los resultados de la prueba de factores específicos de la selección particular de reactivos? Si otro investigador que trabaja de manera independiente prepara otra prueba de acuerdo con las mismas especificaciones, ¿qué tanto diferirá la puntuación que el mismo individuo obtenga en ambas pruebas?

Supongamos que se ha elaborado una prueba de vocabulario de 40 reactivos como medida de la comprensión verbal general. Digamos también que con el mismo propósito se prepara otra lista de 40 palabras nuevas y que los reactivos se redactan con igual cuidado para cubrir el mismo grado de dificultad que la primera prueba. Las diferencias en los resultados obtenidos por los mismos individuos en esas dos pruebas ilustran la varianza de error considerada. Debido a factores fortuitos de las experiencias de individuos distintos, la dificultad relativa de las dos listas diferirá en cierto grado de una persona a otra. Así, para el individuo A la primera lista puede contener un mayor número de palabras desconocidas que la segunda, que, por su parte, puede abarcar un número muy grande de palabras que ignora el individuo B. Aun cuando el conocimiento general de las palabras (es decir, las "puntuaciones verdaderas") de ambos individuos es aproximadamente igual, B superará a A en la primera lista mientras que A aventajará a B en la segunda; por lo tanto, la posición relativa de los dos será inversa en las dos listas debido a las diferencias fortuitas en la selección de los reactivos.

Como la confiabilidad test-retest, la confiabilidad de formas alternas siempre debe ser acompañada por el informe de la duración del intervalo entre las aplicaciones de la prueba y de una descripción de las experiencias intermedias relevantes. Si las dos formas se aplican *en sucesión inmediata*, la correlación muestra la confiabilidad únicamente entre las formas, no entre las ocasiones. En este caso, la varianza de error representa fluctuaciones en el desempeño de un conjunto de reactivos a otro, pero no fluctuaciones temporales.

Es evidente que en la preparación de formas alternas deben tomarse las precauciones necesarias para asegurar que en realidad son paralelas. En principio, deben hacerse independientemente y diseñarse de modo que cumplan las mismas especificaciones; han de contener el mismo número de reactivos, y éstos tienen que expresarse de la misma forma, cubrir el mismo contenido y ser de igual alcance y nivel de dificultad. Además, es necesario supervisar la equivalencia de las instrucciones, límites de tiempo, ejemplos ilustrativos, formato y todos los otros aspectos de la prueba.

Disponer de formas de prueba paralelas no sólo es deseable por su conveniencia para la determinación de la confiabilidad de la prueba, puesto que también son útiles en los estudios de seguimiento o para investigar los efectos de algún factor experimental que interviene en el desempeño de la prueba. El uso de varias formas alternas brinda asimismo un medio para reducir la posibilidad de preparación o de fraudes.

Aunque de mayor aplicación que la confiabilidad test-retest, la confiabilidad de formas alternas también tiene ciertas limitaciones. En primer lugar, si las funciones de conducta consideradas son muy susceptibles al efecto de la práctica el uso de formas alternas reduce pero no elimina dicho efecto. Es obvio que si todos los examinados muestran la misma mejora con la repetición, la correlación entre sus resultados no se altera porque añadir una cantidad constante a cada puntuación no modifica el coeficiente de correlación; no obstante, es mucho más probable que los individuos difieran en el grado de su mejora debido a su experiencia con materiales similares, a la motivación para presentar la prueba y a otros factores. En tales condiciones, el efecto de la práctica representa otra fuente de varianza que tenderá a reducir la correlación entre las dos formas. Si este efecto es pequeño, la reducción será insignificante.

Otra cuestión relacionada atañe al grado en que la naturaleza de la prueba cambia con la repetición; por ejemplo, en ciertos problemas de ingenio, una vez que ha encontrado la solución del primero, la mayoría de la gente puede resolver fácilmente cualquier reactivo que siga el mismo principio. En este caso, no basta con cambiar el contenido de los reactivos de la segunda forma para eliminar este acarreo de la primera. Por último, dadas las dificultades prácticas para elaborar formas verdaderamente equivalentes, muchas pruebas no disponen de formas alternas. Por todo lo anterior, a menudo se requieren otras técnicas para estimar la confiabilidad de la prueba.

Confiabilidad de división por mitades. Mediante diversos procedimientos de división por mitades es posible obtener una medida de confiabilidad a partir de una única aplicación de una forma de la prueba, ya que al dividirla en mitades equivalentes se obtienen dos puntuaciones de cada persona. Es obvio que la confiabilidad de división por mitades proporciona una medida de la consistencia del contenido muestreado, pero no de la estabilidad temporal de las puntuaciones porque el procedimiento consiste en una única sesión de prueba. Este tipo de coeficiente de confiabilidad se conoce como coeficiente de consistencia interna, ya que se requiere una sola aplicación de una única forma.

Puesto que cualquier prueba puede dividirse de muchas maneras distintas, el primer problema es cómo dividirla para obtener las mitades más equivalentes para encontrar la confiabilidad de división por mitades. En la mayor parte de los casos, es posible que las mitades no sean equivalentes por diferencias en la naturaleza y el nivel de dificultad de los reactivos, por los efectos acumulativos de calentamiento, práctica, fatiga, aburrimiento y por cualquier otro factor que varíe progresivamente desde el inicio hasta el final de la prueba. Un procedimiento adecuado para casi todos los propósitos es separar las puntuaciones de los reactivos nones y pares. Si los reactivos están dispuestos en un orden aproximado de dificultad, la división producirá calificaciones casi equivalentes de ambas mitades. Una precaución que debe tomarse al hacer la división entre nones y pares tiene que ver con los grupos de reactivos que abordan un solo problema, como las preguntas que se refieren a un diagrama mecánico particular o a cierto pasaje en una prueba de lectura. En este caso, el grupo entero de reactivos debe asignarse intacto a una u otra mitad. Si los reactivos de dicho grupo se colocaran en mitades diferentes, la similitud de los resultados estaría inflada de manera espuria,

porque cualquier error en la comprensión del problema afectaría a los reactivos de ambas mitades.[3]

Una vez que se tienen las puntuaciones de ambas mitades de cada sujeto, pueden correlacionarse por el método usual, pero no pierda de vista que esta correlación sólo proporciona la confiabilidad de media prueba; por ejemplo, si la prueba consta de 100 reactivos, la correlación se calcula entre dos conjuntos de calificaciones, cada uno basado únicamente en 50 reactivos. En contraste, tanto en la confiabilidad de test-retest como en la de formas alternas cada resultado se basa en el número total de los reactivos de la prueba.

Si todo lo demás se mantiene igual, mientras más extensa sea la prueba más confiable resultará.[4] Es razonable esperar que con una muestra mayor de conducta pueda obtenerse una medida más adecuada y consistente. El efecto que tendrá alargar o acortar la prueba sobre su coeficiente puede estimarse por medio de la fórmula de Spearman-Brown que se presenta a continuación:

$$r_{nn} = \frac{nr_{tt}}{1 + (n-1)r_{tt}}$$

en la que r_{nn} es el coeficiente estimado, r_{tt} el coeficiente obtenido y n el número de veces que la prueba se alarga o se acorta. Así, si el número de reactivos del instrumento aumenta de 25 a 100, n es igual a cuatro; si disminuye de 60 a 30, n es igual a $^1/_2$. La fórmula de Spearman-Brown se emplea mucho para determinar la confiabilidad por el método de división por mitades, y son muchos los manuales de pruebas que informan de este modo la confiabilidad. Cuando se aplica a la confiabilidad de división por mitades, la fórmula requiere que se duplique el tamaño de la prueba. En esas condiciones, puede simplificarse de la siguiente manera:

$$r_{tt} = \frac{2r_{mm}}{1 + r_{mm}}$$

en la que r_{mm} es la correlación de ambas mitades.

Rulon (1939) elaboró un método alternativo para encontrar la confiabilidad de división por mitades, el cual únicamente requiere la varianza de las *diferencias* entre las puntuaciones de cada individuo en las dos mitades de la prueba (DE_d^2) y la varianza de las puntuaciones totales (DE_x^2); estos dos valores se incluyen en la siguiente fórmula, que arroja directamente la confiabilidad de toda la prueba:

$$r_{tt} = 1 - \frac{DE_d^2}{DE_x^2}$$

Es interesante observar la relación de esta fórmula con la definición de la varianza de error. Cualquier diferencia entre las puntuaciones de una persona en las dos mitades de la prueba representa una varianza de error. La varianza de esas diferencias, dividida

[3] Ahora hay suficientes evidencias empíricas que apoyan esta expectativa y se ha acumulado una cantidad considerable de investigaciones sobre el tratamiento estadístico de esos agrupamientos integrados de reactivos o "testlets" (Sireci, Thissen y Wainer, 1991).

[4] Sin embargo, alargar una prueba incrementará su consistencia sólo en términos del muestreo de contenido, pero no su estabilidad con el tiempo (véase Cureton, 1965; Cureton *et al.*, 1973).

entre la varianza de las puntuaciones totales, da la proporción de varianza de error en los resultados que, cuando esta varianza de error se resta de 1.00, se obtiene la proporción de varianza "verdadera" para un uso específico de la prueba, que es igual al coeficiente de confiabilidad.

Confiabilidad de Kuder-Richardson y coeficiente alfa. El cuarto método para encontrar la confiabilidad, que también requiere una sola aplicación de una única forma, se basa en la consistencia de las puntuaciones a todos los reactivos de la prueba. Esta *consistencia entre reactivos* está influida por dos fuentes de varianza de error: (1) el muestreo de contenido (como en la confiabilidad de formas alternas y de división por mitades), y (2) la heterogeneidad del área de conducta muestreada. Entre más homogénea sea el área mayor será la consistencia entre reactivos; por ejemplo, si una prueba sólo incluye multiplicaciones mientras que otra abarca sumas, restas, multiplicaciones y divisiones, es probable que la primera prueba muestre mayor consistencia entre reactivos que la segunda, en la que, como es más heterogénea, el examinado puede hacer un mejor trabajo en las restas que en las otras operaciones aritméticas; otro puede obtener mejores resultados en las divisiones, pero menos en las sumas, restas y multiplicaciones, etc. Un ejemplo más extremo sería el caso de una prueba constituida por 40 reactivos de vocabulario comparada con otra que tiene 10 reactivos de vocabulario, 10 de relaciones espaciales, 10 de razonamiento aritmético y 10 de rapidez perceptual. En esta última puede esperarse poca o ninguna relación entre el desempeño de un individuo en los diferentes tipos de reactivos.

Es evidente que las puntuaciones de la prueba serán menos ambiguas cuando se deriven de instrumentos relativamente homogéneos. Suponga que en la prueba muy heterogénea de 40 reactivos, Pérez y Jiménez obtienen un resultado de 20. ¿Podemos concluir que el desempeño de ambos en la prueba fue igual? De ninguna manera, pues Pérez pudo haber resuelto correctamente los 10 reactivos de vocabulario, los 10 de rapidez perceptual y ninguno de los reactivos de razonamiento aritmético y de relaciones espaciales. Por su parte, Jiménez pudo haber recibido la calificación de 20 por la correcta solución de cinco reactivos de rapidez perceptual, cinco de relaciones espaciales, 10 de razonamiento aritmético y ninguno de vocabulario.

Por supuesto, hay muchas otras combinaciones que producen el mismo resultado total de 20, que tendría un significado muy diferente al obtenerse de tan distintas organizaciones de reactivos. Por otro lado, en la relativamente homogénea prueba de vocabulario, si todos los reactivos se dispusieron en orden ascendente de dificultad, es posible que una puntuación de 20 signifique que el examinado ha tenido éxito aproximadamente en las primeras 20 palabras. Tal vez falló en dos o tres de las más sencillas y respondió bien a dos o tres de las más difíciles después a la vigésima, pero estas variaciones individuales son ligeras en comparación con las que se encuentran en una prueba más heterogénea.

A este respecto, una cuestión de gran importancia es si el criterio que la prueba pretende predecir es en sí mismo homogéneo o heterogéneo. Aunque se prefieren las pruebas homogéneas porque sus resultados permiten una interpretación poco ambigua, es evidente que un solo instrumento homogéneo no es un predictor adecuado de un criterio muy heterogéneo. Más aún, en la predicción de éste, la heterogeneidad de

los reactivos no representa por fuerza una varianza de error. Los tests tradicionales de inteligencia son un buen ejemplo de instrumentos heterogéneos diseñados para predecir criterios heterogéneos; sin embargo, en un caso así puede ser deseable formular varias pruebas relativamente homogéneas, de las que cada una mida una fase diferente del criterio heterogéneo. De este modo puede combinarse una interpretación no ambigua de los resultados de la prueba con una adecuada cobertura del criterio.

El procedimiento más común para encontrar la consistencia entre reactivos se debe a Kuder y Richardson (1937). Como en los métodos de división por mitades, la consistencia entre reactivos se encuentra en una sola aplicación de una única prueba; sin embargo, en lugar de requerir las puntuaciones de las dos mitades, la técnica se basa en el examen del desempeño en cada reactivo. De las diversas fórmulas derivadas del artículo original, la de mayor aplicación, conocida comúnmente como la "fórmula 20 de Kuder-Richardson", es la siguiente:

$$r_{tt} = \left(\frac{n}{n-1}\right)\frac{DE_t^2 - \Sigma pq}{DE_t^2}$$

en la que r_{tt} es el coeficiente de confiabilidad de toda la prueba, n el número de reactivos y DE_t la desviación estándar de la puntuación total de la prueba. El único término nuevo de la fórmula, Σpq, se encuentra calculando la proporción de personas que aciertan (p) y la de personas que no aciertan (q) en cada reactivo. Para obtener Σpq se calcula para cada reactivo el producto de p y q, y luego se suman los productos de todos los reactivos. Este método requiere de pocos cálculos adicionales porque al elaborar el instrumento rutinariamente se registra p para encontrar el grado de dificultad de cada reactivo.

Puede demostrarse matemáticamente que el coeficiente de confiabilidad de Kuder-Richardson es en realidad la media de todos los coeficientes de división por mitades que resultan de las diferentes divisiones de una prueba (Cronbach, 1951).[5] Por otro lado, el coeficiente de división por mitades en general se basa en una división planeada para obtener conjuntos equivalentes de reactivos; por lo tanto, a menos que los reactivos de la prueba sean muy homogéneos, el coeficiente de Kuder-Richardson será menor que la confiabilidad de división por mitades. Un ejemplo extremo permitirá ilustrar la diferencia. Suponga que preparamos una prueba de 50 reactivos de 25 tipos diferentes, de modo que los reactivos 1 y 2 son de vocabulario, 3 y 4 de razonamiento aritmético, 5 y 6 de orientación espacial, etc. En teoría, habrá una relación tan estrecha entre las puntuaciones de los reactivos nones y pares, que arrojaría un elevado coeficiente de confiabilidad de división por mitades; sin embargo, la homogeneidad de la prueba sería muy baja ya que habría poca consistencia en el desempeño del conjunto completo de 50 reactivos. En este ejemplo, se esperaría que la confiabilidad de Kuder-Richardson fuera mucho menor que la confiabilidad de división por mitades. En efecto, la diferencia entre ambos coeficientes de confiabilidad puede servir como un indicador grueso de la heterogeneidad de la prueba.

[5] Esto es estrictamente cierto sólo cuando los coeficientes de división por mitades se obtienen con la fórmula de Rulon (basada en la varianza de las diferencias entre las puntuaciones de ambas mitades), pero no cuando se obtienen con la correlación de mitades y la fórmula de Spearman-Brown (Novick y Lewis, 1967).

La fórmula de Kuder-Richardson puede aplicarse a pruebas cuyos reactivos se califiquen como aciertos o errores, o de acuerdo con algún otro sistema de todo o nada. Ahora bien, los reactivos de ciertas pruebas pueden obtener diferentes puntuaciones; por ejemplo, en un inventario de personalidad el sujeto puede recibir en un reactivo una calificación numérica diferente dependiendo de si verifica sus respuestas "de manera regular", "a veces", "rara vez" o "nunca". Para estos casos puede derivarse una fórmula generalizada conocida como coeficiente alfa (Cronbach, 1951; Kaiser y Michael, 1975; Novick y Lewis, 1967), en la que $\Sigma\ (DE_i^2)$, la suma de las varianzas de las puntuaciones del reactivo, sustituye a Σpq. El procedimiento consiste en encontrar la varianza de todas las puntuaciones individuales de cada reactivo y sumar las varianzas de todos los reactivos. La fórmula completa del coeficiente alfa es:

$$r_{tt} = \left(\frac{n}{n-1}\right)\frac{DE_t^2 - \Sigma(DE_i^2)}{DE_t^2}$$

Confiabilidad entre calificadores. Ahora debe ser evidente que las distintas formas de confiabilidad difieren en los factores que incluyen bajo la varianza de error. En un caso, ésta cubre las fluctuaciones temporales, en otro se refiere a la diferencia entre conjuntos de reactivos paralelos y en otro más comprende cualquier inconsistencia entre reactivos. Por otra parte, los factores *excluidos* de las medidas de la varianza de error son, de manera general, de dos clases: (a) los factores cuya varianza debe permanecer en las puntuaciones porque forman parte de las diferencias consideradas, y (b) los factores irrelevantes que pueden controlarse experimentalmente. Por ejemplo, no se acostumbra informar el error de medición que resulta cuando una prueba se aplica en condiciones de distracción o con un tiempo límite mayor o menor a lo especificado en el manual. Los errores de cronometración y las distracciones serias pueden eliminarse de la situación de prueba, por lo que no es necesario manifestar coeficientes de confiabilidad especiales que correspondan a la "varianza de distracción" o a la "varianza de cronometración".

Además, la mayor parte de las pruebas proporciona procedimientos tan estandarizados para la administración y calificación que la varianza de error atribuible a esos factores es insignificante. Esto es en especial cierto para el grupo de pruebas diseñadas para aplicarse colectivamente y para ser calificadas por computadora. Con tales instrumentos, sólo es necesario seguir minuciosamente los procedimientos prescritos y supervisarlos con cuidado. Por otro lado, en el caso de los instrumentos clínicos empleados en exámenes intensivos individuales hay evidencias de una considerable *varianza del examinador*. Mediante diseños experimentales especiales es posible separar esta varianza de la que puede atribuirse a las fluctuaciones temporales en la condición del examinado o al uso de formas alternas de la prueba.

Una fuente de varianza de error que puede supervisarse de manera sencilla es la *varianza del calificador*. Algunas categorías de pruebas (sobre todo las de creatividad y las proyectivas de personalidad) dejan mucho al juicio del calificador. En el caso de estas pruebas hay tanta necesidad de una medida de confiabilidad del calificador como de los coeficientes más comunes de confiabilidad. La confiabilidad del calificador se consigue con una muestra de pruebas calificadas independientemente por dos o mas examinadores. Los dos resultados se correlacionan de acuerdo con la forma común y el coeficiente de correlación es la medida de la confiabilidad del calificador. Esta clase de

confiabilidad suele calcularse cuando se emplean en la investigación instrumentos de calificación subjetiva. Los manuales deben informar en qué casos es apropiada.

Recapitulación. Las diferentes clases de coeficientes de confiabilidad revisados en esta sección se resumen en las tablas 4.2 y 4.3. En la primera se clasifican las operaciones seguidas para obtener cada tipo de confiabilidad en relación con el número de formas de prueba y de sesiones de examinación que se requieren. La tabla 4.3 muestra las fuentes de varianza tratadas como varianza de error por cada procedimiento.

Cualquier coeficiente de confiabilidad puede interpretarse directamente en términos del *porcentaje de la varianza de la calificación* que puede atribuirse a diferentes fuentes. Así, un coeficiente de confiabilidad de .85 significa que 85 por ciento de la varianza en las puntuaciones de la prueba depende de la varianza verdadera en el rasgo medido y que 15 por ciento depende de la varianza de error (según lo haya definido operacionalmente el procedimiento específico). El lector con conocimientos de estadística recordará que el *cuadrado* del coeficiente de correlación representa la proporción de la varianza común. En realidad, la proporción de varianza verdadera de las puntuaciones de una prueba es el cuadrado de la correlación entre las puntuaciones obtenidas en una sola forma de la prueba y las puntuaciones verdaderas, que están libres de errores fortuitos. Esta correlación, conocida como índice de confiabilidad,[6] es igual a la raíz cuadrada del coeficiente de confiabilidad ($\sqrt{r_{tt}}$). Cuando el índice de confiabilidad a su vez se eleva al cuadrado, el resultado es el coeficiente de confiabilidad (r_{tt}), que puede entonces interpretarse directamente como el porcentaje de varianza verdadera para el uso al que se destinó la prueba.

Los diseños experimentales que producen más de un tipo de coeficiente de confiabilidad para el mismo grupo permiten el análisis de varianza total en distintos componentes. Consideremos el siguiente ejemplo hipotético. Se aplicaron las formas A y B de una prueba de creatividad con un intervalo de dos meses a 100 niños de sexto grado. La confiabilidad de formas alternas resultó ser de .70. También puede calcularse un coefi-

Tabla 4.2

Técnicas para medir la confiabilidad en relación con la forma de la prueba y la sesión de aplicación

Sesiones de aplicación requerida	Formas de prueba requeridas	
	Una	*Dos*
Una	División por mitades Kuder-Richardson	Formas alternas (Inmediata)
Dos	Test-Retest	Formas alternas (Demorada)

[6] Gulliksen (1950, capítulos 2 y 3) ofrece las derivaciones del índice de confiabilidad basadas en dos diferentes conjuntos de suposiciones.

Tabla 4.3

Fuentes de varianza de error en relación a los coeficientes de confiabilidad

Tipo de coeficiente de confiabilidad	Varianza de error
Test-Retest	Muestreo de tiempo
Formas alternas (inmediata)	Muestreo de contenido
Formas alternas (demorada)	Muestreo de tiempo y muestreo de contenido
División por mitades	Muestreo de contenido
Kuder-Richardson y Coeficiente alfa	Muestreo de contenido y heterogeneidad de contenido
Del calificador	Diferencias entre calificadores

ciente de confiabilidad[7] de división por mitades para las respuestas a cualquiera de las formas. Este coeficiente, obtenido por la fórmula Spearman-Brown, es de .80. Finalmente, al segundo calificador se le asignó una muestra aleatoria de 50 pruebas, de lo que se obtuvo una confiabilidad entre calificadores de .92. Los tres coeficientes de confiabilidad pueden ahora analizarse para llegar a las varianzas de error que muestran la tabla 4.4 y la figura 4.4. Observe que al restar la varianza de error atribuible únicamente al muestreo de contenido (confiabilidad de división por mitades) de la varianza de error atribuible al muestreo de contenido y de tiempo (confiabilidad de formas alternas), encontramos que .10 de la varianza puede atribuirse sólo al muestreo de tiempo. Al sumar las varianzas de error atribuibles al muestreo de contenido (.20), el muestreo de tiempo (.10) y la diferencia entre calificadores (.08) se obtiene una varianza de error total de .38 y, por ende, una varianza verdadera de .62. Las proporciones, expresadas en

Tabla 4.4

Análisis de las fuentes de varianza de error en una prueba hipotética

De la confiabilidad de formas alternas demoradas:	$1 - .70 = .30$	(muestreo de tiempo más muestreo de contenido)
De división por mitades, confiabilidad de Spearman-Brown:	$1 - .80 = .20^*$	(muestreo de contenido)
Diferencia	$.10^*$	(muestreo de tiempo)
De la confiabilidad del calificador:	$1 - .92 = .08^*$	(diferencia entre calificadores)

Medida total de la varianza de error* = .20 + .10 + .08 = .38*
Varianza verdadera = 1 − .38 = .62

* Varianza de error

[7] Para una mejor estimación del coeficiente de consistencia interna, pueden calcularse correlaciones de división por mitades para cada forma y promediarse los dos coeficientes por los procedimientos estadísticos apropiados (por ejemplo, una transformación-z de Fisher).

los términos más familiares de porcentajes, se muestran gráficamente en la figura 4.4. Esta distribución de las fuentes de varianza es la esencia de la llamada teoría de la generalización de la confiabilidad. Los diseños experimentales complejos que permiten la evaluación simultánea de más fuentes de varianza de la puntuación y las interacciones entre ellas pueden encontrarse en otros tratamientos del tema, como en Brennan (1984), Cronbach *et al.* (1972), Feldt y Brennan (1989), y Shavelson y Webb (1991).

CONFIABILIDAD DE LAS PRUEBAS DE VELOCIDAD

Tanto en la elaboración de la prueba como en la interpretación de las puntuaciones, es importante distinguir entre las mediciones de velocidad y las de poder. Una *prueba de velocidad* es aquella en la que las diferencias individuales dependen por completo de la rapidez de la ejecución. Dicha prueba consta de reactivos de dificultad uniformemente baja, todos están dentro del nivel de habilidad de las personas para las cuales están destinados. El límite de tiempo es tan reducido que nadie puede terminar todos los reactivos. En esas condiciones, el resultado de cada sujeto sólo refleja la velocidad con la que trabajó. Por otro lado, una *prueba de poder* tiene un límite de tiempo lo suficientemente amplio como para permitir que cualquiera intente resolver todos los reactivos. La dificultad de éstos está graduada, y el instrumento incluye algunos muy difíciles, por lo que nadie puede obtener una puntuación perfecta.

Advierta que tanto las pruebas de velocidad como las de poder están diseñadas para impedir las puntuaciones perfectas. La razón de tal precaución es que este tipo de resultados es indeterminado, ya que es imposible saber qué tan elevada habría sido la puntuación del individuo si se hubieran incluido más reactivos o reactivos más difíciles. Para que cada individuo muestre lo que es capaz de lograr, la prueba debe proporcionar un tope adecuado, ya sea en número de reactivos o en el nivel de dificultad. Una excepción a esta regla se encuentra en las pruebas de destreza, como las referidas a dominio que revisamos en el capítulo 3. Su propósito no es establecer los límites de lo que puede hacer el individuo, sino determinar si ha alcanzado o no el nivel preestablecido de ejecución.

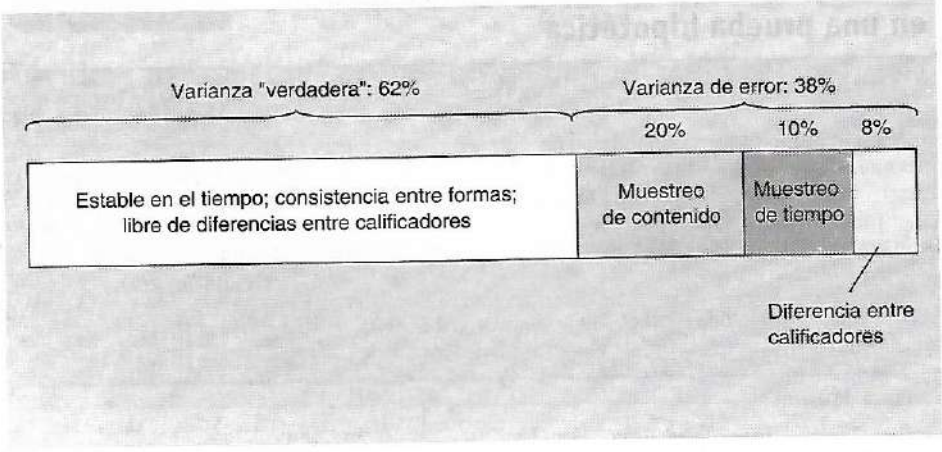


Figura 4.4. Distribución del porcentaje de varianza de la puntuación de una prueba hipotética.

En la práctica, la distinción entre las pruebas de velocidad y las de poder es de grado, ya que la mayor parte depende en diversas proporciones tanto del poder como de la velocidad. Para cada prueba se requiere de información sobre estas proporciones, no sólo para comprender qué es lo que mide el instrumento, sino también para elegir los procedimientos apropiados para evaluar su confiabilidad. Los coeficientes de confiabilidad de un solo ensayo, como los que se obtienen con las técnicas de pares-nones o de Kuder-Richardson, no son adecuados para las pruebas de velocidad, pues en la medida en que las diferencias individuales de las puntuaciones dependan de la velocidad de la ejecución, los coeficientes de confiabilidad que señalan esos métodos estarán elevados en forma espuria. Un ejemplo extremo ayudará a aclarar ese punto. Supongamos que una prueba de 50 reactivos depende por completo de la velocidad, por lo que las diferencias individuales en las puntuaciones dependen por completo del número de reactivos contestados más que de los errores. Entonces, si el individuo A obtiene una calificación de 44, obviamente acertó en 22 reactivos nones y en 22 reactivos pares. De modo similar, el individuo B, con una puntuación de 34, acertó en 17 reactivos nones y 17 pares. En consecuencia, salvo por algunos errores accidentales insignificantes en algunos reactivos, la correlación entre las puntuaciones nones y pares será perfecta, o de +1.00; pero resulta del todo falsa y no brinda información sobre la confiabilidad de la prueba.

Una evaluación de los procedimientos seguidos para encontrar la confiabilidad de división por mitades y la de Kuder-Richardson demostrará que ambos se basan en la consistencia en el *número de errores* cometido por el examinado. Ahora bien, si las diferencias individuales en los resultados de la prueba no dependen de los errores, sino de la velocidad, es evidente que la medida de confiabilidad debe basarse en la consistencia de la *velocidad del trabajo*. Cuando la ejecución en la prueba depende de una combinación de velocidad y poder, el coeficiente de confiabilidad de una sola aplicación será menor a 1.00, pero continuará siendo elevado en forma errónea. En la medida en que la velocidad influya de modo apreciable en las diferencias individuales de las puntuaciones, los coeficientes de confiabilidad de una sola aplicación no podrán interpretarse adecuadamente.

¿De qué otros procedimientos se dispone para determinar la confiabilidad de pruebas de velocidad? De ser aplicables, es factible emplear la técnica test-retest o la de formas equivalentes para evaluar la confiabilidad de las pruebas de velocidad. También pueden emplearse las técnicas de división por mitades, siempre que la división se haga en términos de tiempo más que de reactivos. En otras palabras, las puntuaciones deben basarse en partes de la prueba cronometradas por separado. Una forma de efectuar dicha división consiste en aplicar dos mitades equivalentes de la prueba con distintos límites de tiempo; por ejemplo, se imprimen los reactivos nones y los pares en hojas separadas y se asigna a cada conjunto la mitad del límite de tiempo de la prueba entera. Ese procedimiento es equiparable a aplicar dos formas equivalentes de la prueba en sucesión inmediata; sin embargo, cada forma tiene la mitad de la extensión de la prueba, mientras que las puntuaciones del examinado normalmente se basan en la prueba entera. Es por ello que debe utilizarse la fórmula de Spearman-Brown, o alguna otra apropiada, para encontrar la confiabilidad de la prueba entera.

Si no es posible aplicar por separado las dos mitades de la prueba, un procedimiento alternativo consiste en dividir el tiempo total en cuartos y encontrar una calificación para cada uno. Esto se logra con facilidad al hacer que los examinados marquen el reactivo en el que están trabajando cada vez que el examinador da una señal conve-

nida. Entonces, se combina el número de reactivos resueltos correctamente dentro del primer y del cuarto cuartos para representar el resultado de una mitad, mientras que los del segundo y tercer cuartos sirven para la puntuación de la otra mitad. Esta combinación de cuartos tiende a equilibrar los efectos acumulativos de la práctica, la fatiga y otros factores, y este método es satisfactorio cuando el grado de dificultad de los reactivos no está graduado progresivamente.

¿Cuándo una prueba es apreciablemente rápida? ¿En qué condiciones deben observarse las precauciones especiales que indicamos en esta sección? Es evidente que el mero empleo de un límite de tiempo no significa que se trate de una prueba de velocidad. Si todos los examinados terminan dentro del límite de tiempo concedido, la rapidez en la ejecución de la tarea no participa en la determinación de las puntuaciones. El porcentaje de los que no logran completar la prueba debe tomarse como un indicador crudo de velocidad contra poder; con todo, incluso si nadie termina la prueba el papel de la velocidad puede ser insignificante; por ejemplo, si todos completan exactamente 40 reactivos de una prueba de 50, las diferencias individuales relativas a la velocidad están totalmente ausentes, aunque nadie tuviera tiempo para tratar de resolver todos los reactivos.

Por supuesto, la pregunta crucial es: "¿en qué medida pueden atribuirse a la velocidad las diferencias individuales en los resultados de la prueba?" En términos más técnicos, queremos saber qué proporción de la varianza total de las puntuaciones es varianza de velocidad. Para estimar en forma aproximada esta proporción, se encuentra la varianza del número de reactivos completados por diferentes personas y se divide entre la varianza total de las puntuaciones de la prueba (DE_c^2/DE_t^2). En el ejemplo del párrafo anterior, en el que cada individuo termina 40 reactivos, el numerador de la fracción sería de cero, ya que no hay diferencias individuales en el número de reactivos completados ($DE_c^2 = 0$); y en una prueba de poder el índice entero también sería igual a cero. Por otro lado, si la varianza total de la prueba (DE_t^2) es atribuible a diferencias individuales en la velocidad, las dos varianzas serán la misma y la razón será igual a 1.00. Se han concebido otros procedimientos más refinados para determinar esta proporción, pero su estudio detallado escapa al alcance de este libro.

Un ejemplo del efecto de la velocidad sobre los coeficientes de confiabilidad de un solo ensayo se encuentra en los datos de una investigación de la primera edición de la prueba SRA del Test de Habilidades Mentales Primarias para las edades de 11 a 17 años (Anastasi y Drake, 1954). En este estudio, primero se determinó la confiabilidad de cada prueba por el procedimiento común de pares-nones (esos coeficientes se presentan en el primer renglón de la tabla 4.5). Luego se correlacionaron las puntuaciones de las mitades cronometradas por separado para calcular los coeficientes de confiabilidad (que se presentan en el segundo renglón de la tabla 4.5). El cálculo de los índices de velocidad demostró que la prueba de significado verbal era principalmente una prueba de poder, mientras que la prueba de razonamiento dependía algo más de la velocidad. Las pruebas de espacio y número mostraron que dependen en gran medida de la velocidad. En la tabla 4.5 puede observarse que, cuando se calcula apropiadamente, la confiabilidad de la prueba de espacio es .75, en contraste con un coeficiente de pares-nones de .90, elevado en forma incorrecta. De modo similar, la confiabilidad de la prueba de razonamiento cae de .96 a .87, y la de la prueba de número cae de .92 a .83. Por otro lado, la confiabilidad de la prueba de significado verbal, que de algún modo no es de rapidez, muestra una diferencia insignificante cuando se calcula por ambos métodos.

Tabla 4.5

Coeficientes de confiabilidad de cuatro de los Tests SRA de Habilidades Mentales Primarias para las edades de 11 a 17 (primera edición)

Coeficiente de confiabilidad encontrado por	Significado verbal	Razonamiento	Espacio	Número
Método de pares-nones de un ensayo	.94	.96	.90	.92
Mitades cronometradas por separado	.90	.87	.75	.83

(Datos de Anastasi y Drake, 1954.)

DEPENDENCIA DE LOS COEFICIENTES DE CONFIABILIDAD DE LA MUESTRA EXAMINADA

Variabilidad. Una condición importante que afecta el tamaño del coeficiente de confiabilidad es la naturaleza del grupo en el que fue medido. En primer lugar, el grado de las diferencias individuales del grupo influye en cualquier coeficiente de correlación. Si la habilidad ortográfica de cada miembro de un grupo fuera muy parecida, la correlación de la ortografía con cualquier otra habilidad estaría cerca de cero, por lo que en ese grupo resultaría imposible predecir la posición de un individuo en cualquier otra habilidad a partir del conocimiento de su puntuación en ortografía.

Otro ejemplo menos extremo se encuentra en la correlación entre dos tests de aptitud, como uno de comprensión verbal y otro de razonamiento aritmético. Si se aplicaran estos instrumentos a una muestra muy homogénea (por ejemplo a un grupo de 300 universitarios de segundo año), la correlación entre ambos tests probablemente sería muy baja. Debido a la restricción del rango, en esta selecta muestra de estudiantes universitarios hay poca relación entre la habilidad verbal de cualquier individuo y su habilidad de razonamiento numérico. Por otro lado, si las pruebas se aplicaran a una muestra heterogénea de 300 personas (que comprenda desde individuos con retraso mental hasta universitarios) sin duda se obtendría una elevada correlación entre las dos pruebas. Los retardados obtendrían calificaciones menores que los universitarios en ambas pruebas, y se encontrarían relaciones similares para otros subgrupos dentro de esta muestra tan heterogénea.

El análisis del diagrama de dispersión hipotético de la figura 4.5 es otro ejemplo de la forma en que los coeficientes de correlación dependen de la variabilidad, o del grado de diferencias individuales, dentro del grupo. Este diagrama de dispersión muestra una elevada correlación positiva en todo el grupo heterogéneo, ya que las entradas se agrupan alrededor de la diagonal que se extiende del extremo inferior izquierdo al extremo superior derecho. Pero si consideramos sólo al subgrupo que cae dentro del pequeño rectángulo en la porción superior derecha del diagrama, resulta evidente que la correlación entre las dos variables es cercana a cero. Los individuos que caen dentro de este rango restringido en ambas variables representan un grupo muy homogéneo, como es el caso de los universitarios de segundo año.

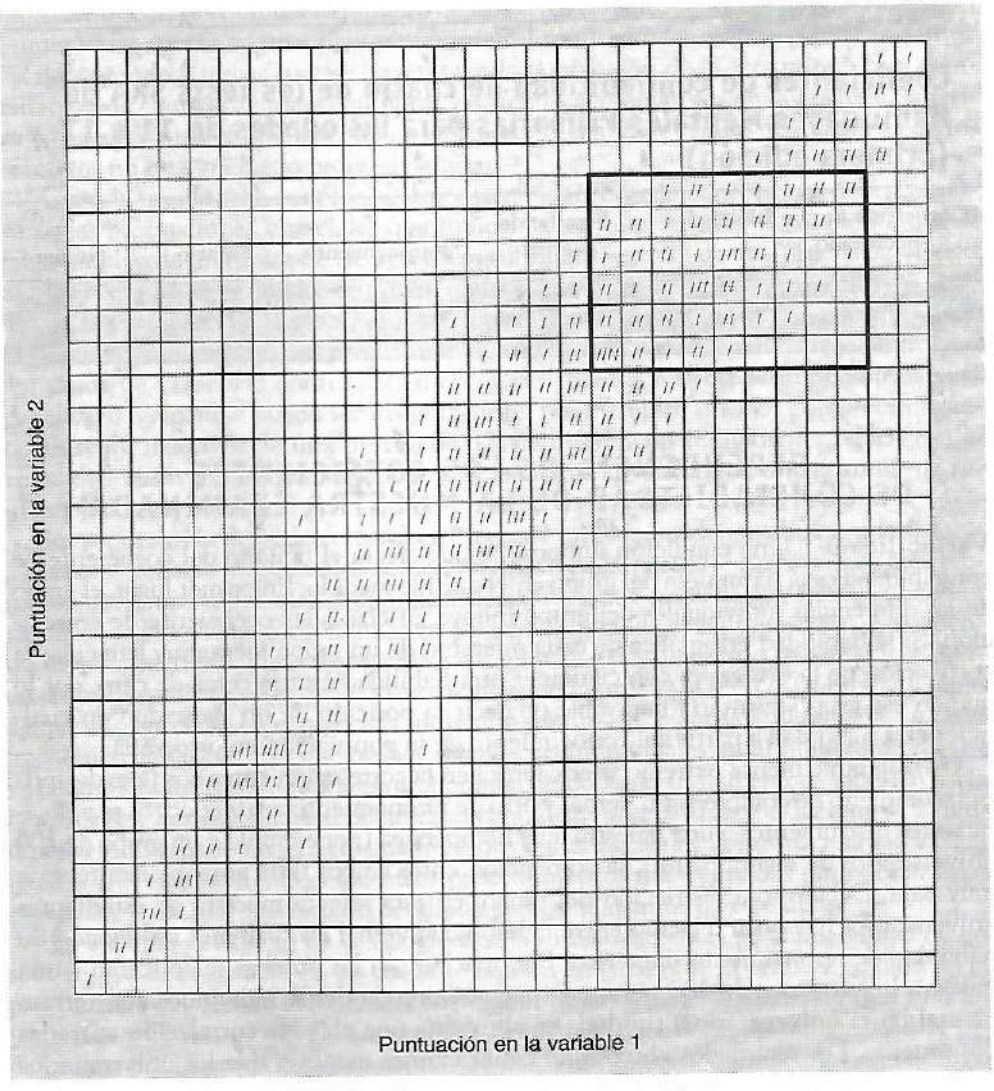


Figura 4.5. El efecto de un rango restringido sobre el coeficiente de correlación.

Al igual que todos los coeficientes, los de confiabilidad dependen de la variabilidad de la muestra en la que se encontraron. Así, si el coeficiente de confiabilidad que indica el manual de una prueba se calculó con un grupo que incluía a niños de cuarto grado a estudiantes de secundaria, no puede suponerse que la confiabilidad será tan elevada con una muestra de, digamos, estudiantes de preparatoria. Cuando pretende usarse una prueba para discriminar diferencias individuales dentro de una muestra más homogénea que el grupo de estandarización, el coeficiente de confiabilidad debe volver a determinarse para esta muestra. Los textos de estadística elemental proporcionan las fórmulas para estimar los coeficientes de confiabilidad esperados cuando la desviación estándar

del grupo aumenta o disminuye; sin embargo, es preferible volver a calcular el coeficiente de confiabilidad con un grupo que sea comparable al que va a resolver la prueba. En el caso de las pruebas diseñadas para cubrir un amplio rango de edad o de habilidad, el manual de la prueba debe informar los coeficientes de confiabilidad por separado para subgrupos relativamente homogéneos dentro de la muestra de estandarización.

Nivel de habilidad. El coeficiente de confiabilidad no sólo varía en la medida en que lo hacen las diferencias individuales en la muestra, sino también entre grupos cuyo nivel promedio de habilidad difiere. Más aún, por lo general esas diferencias no pueden predecirse o estimarse mediante una fórmula estadística, y sólo se descubren aplicando la prueba a grupos que difieren en edad o nivel de habilidad. Esas diferencias en la confiabilidad de una sola prueba pueden deberse en parte al hecho de que distintos niveles de dificultad de la prueba miden una combinación ligeramente diferente de habilidades, o bien a que la longitud de la prueba puede variar con los niveles de edad. Incluso cuando el número de reactivos es el mismo, los extremos superior e inferior pueden incluir un número insuficiente del grado de dificultad apropiado para permitir que los individuos demuestren adecuadamente lo que son capaces de hacer (efectos de tope o de piso). En otras pruebas, la confiabilidad puede ser algo menor para los grupos más jóvenes y menos capaces, ya que sus puntuaciones están influidas por la adivinación.

Como el coeficiente de confiabilidad sólo es aplicable a muestras similares a aquella en la que se calculó, resulta claro que a cada uno debe acompañarlo una descripción detallada del grupo en el que se determinó y es necesario prestar especial atención a la variabilidad y al nivel de habilidad de la muestra. Una práctica deseable que se está extendiendo en la elaboración de las pruebas consiste en fraccionar la muestra de estandarización en subgrupos más homogéneos con respecto a edad, sexo, nivel de educación, ocupación, etc., e informar por separado los coeficientes de confiabilidad para cada subgrupo. En tales condiciones, es más probable que los coeficientes se apliquen a las muestras que en efecto van a realizar la prueba.

ERROR ESTÁNDAR DE MEDICIÓN

Interpretación de las puntuaciones individuales. La confiabilidad de una prueba puede expresarse en términos del error estándar de medición (EEM), conocido también como error estándar de las puntuaciones. Esta medida es en especial adecuada para la interpretación de las puntuaciones individuales, por lo que resulta más útil que el coeficiente de confiabilidad para muchos propósitos de investigación. El error estándar de medición se calcula con facilidad a partir del coeficiente de confiabilidad de la prueba por medio de la siguiente fórmula:

$$\text{EEM} = DE_t\sqrt{1 - r_{tt}}$$

en la que DE_t es la desviación estándar de las puntuaciones de la prueba y r_{tt} es el coeficiente de confiabilidad, ambos calculados en el mismo grupo; por ejemplo, si

el CI de desviación de determinado test de inteligencia tiene una desviación estándar de 15 y un coeficiente de confiabilidad de .89, el EEM del CI en esta prueba es $15\sqrt{1-.89} = 15\sqrt{.11} = 15(.33) = 5$.

Para entender qué nos dice el EEM acerca de una puntuación, supongamos que tenemos un conjunto de 100 CI obtenidos con la prueba anterior por una sola niña, Eugenia. Debido a los errores aleatorios que hemos estudiado en el capítulo, esas puntuaciones varían y caen en una distribución normal alrededor de la verdadera puntuación de Eugenia. La media de esta distribución de 100 puntuaciones puede tomarse como la "puntuación verdadera" para un uso específico de la prueba, y la desviación estándar de la distribución puede considerarse como el EEM. Como cualquier desviación estándar, este error puede interpretarse en términos de la curva normal de frecuencias que vimos en el capítulo 3 (véase la figura 3.3). Recordará que entre la media y $\pm 1\sigma$ se encuentra aproximadamente el 68 por ciento de los casos en una curva normal. Así, podemos concluir que hay más o menos una posibilidad de 2:1 (o 68:32) de que el CI de Eugenia en la prueba fluctúe entre ±1 EEM o cinco puntos a ambos lados de su verdadero CI. Si su verdadero CI es de 110, esperaríamos que obtuviera entre 105 y 115 alrededor de dos terceras partes (68 por ciento).

Si deseamos estar más seguros de nuestra predicción, podemos elegir probabilidades mayores a 2:1. Si volvemos a la figura 3.3 (capítulo 3) veremos que $\pm 3\sigma$ cubre 99.7 por ciento de los casos. Puede verificarse en las tablas de la curva normal de frecuencias que una distancia de 2.58σ a ambos lados de la media comprende exactamente al 99 por ciento de los casos; por consecuencia, hay una probabilidad de 99:1 de que el CI de Eugenia esté dentro de un EEM de 2.58, o (2.58) (5) = 13 puntos, a cualquier lado de su verdadero CI. Podemos entonces afirmar con 99 por ciento de confianza (con una sola posibilidad de error en cada 100) que el CI de Eugenia en una única aplicación de la prueba caerá entre 97 y 123 (110 – 13 y 110 + 13). Si se aplicaran a Eugenia 100 pruebas equivalentes, su CI quedaría una sola vez fuera de este margen de valores.

Por supuesto, en la práctica no obtenemos las puntuaciones verdaderas, sino apenas las obtenidas de una sola aplicación de la prueba. En tales circunstancias, aplicamos el razonamiento anterior en la dirección inversa. Si es poco probable que el resultado de un individuo se desvíe más de 2.58 de EEM de su verdadera calificación, podemos argumentar que su *puntuación verdadera* debe estar dentro de 2.58 de EEM de la puntuación *obtenida*. Aunque no nos es posible asignar una probabilidad a esta afirmación para cualquier puntuación obtenida, *podemos* decir que la afirmación sería correcta para el 99 por ciento de los casos. Con este razonamiento, Gulliksen (1950, pp. 17–20) propuso que el error estándar de medición se utilice, como acabamos de ver, para estimar los límites razonables de la puntuación verdadera de una persona con cualquier resultado obtenido. En términos de dichos "límites razonables" suele interpretarse el error de medición en las pruebas psicológicas, y así lo interpretaremos en este libro.[8]

[8] Se han propuesto otros procedimientos que usan una puntuación "verdadera" estimada como centro del intervalo de confianza (Dudek, 1979; Glutting, McDermott y Stanley, 1987). Si el coeficiente de confiabilidad es elevado, este procedimiento tiene poco efecto; si es bajo, tanto la puntuación verdadera como el tamaño del intervalo de confianza se calculan a partir del mismo falible coeficiente de confiabilidad. Más aún, el procedimiento óptimo varía según el propósito particular para el cual se pretende utilizar las puntuaciones de la prueba (por ejemplo, para hacer una predicción a largo plazo o la evaluación del desempeñó actual).

Desde luego, el error estándar de medición y el coeficiente de confiabilidad son dos formas de expresar la confiabilidad de la prueba. A diferencia del coeficiente de confiabilidad, el error de medición es independiente de la variabilidad del grupo en que se calculó. Expresado en términos de las puntuaciones individuales, permanece sin cambio al encontrarse en un grupo homogéneo o en uno heterogéneo. Por otro lado, al informarse en unidades de calificación, el error de medición no puede compararse directamente de una prueba a otra. Los problemas comunes de la comparabilidad de las unidades surgen entonces cuando los errores de medición se reportan en términos de problemas aritméticos, palabras en una prueba de vocabulario, etc.; por ende, si deseamos comparar la confiabilidad de *diferentes pruebas*, resulta más conveniente el coeficiente de confiabilidad, mientras que el error estándar de medición es más apropiado para interpretar las *puntuaciones individuales*.

Sin embargo, cuando hay mucha variabilidad en los *niveles de habilidad* no puede esperarse que los coeficientes de confiabilidad y los errores de medición permanezcan constantes. Las diferencias en los coeficientes de confiabilidad que revisamos en la sección precedente se mantienen al calcular los errores de medición a diferentes niveles de la misma prueba. Las técnicas TRI de análisis de reactivos citadas en el capítulo 3 ofrecen una solución al problema. Al abarcar un amplio rango de habilidades, esas técnicas brindan una forma de expresar la exactitud de la medición de una prueba en función del nivel de habilidad. El procedimiento arroja una *curva de información de la prueba* que depende únicamente de los reactivos incluidos y permite una estimación del error de medición en cada nivel de habilidad. En el capítulo 7 veremos más a fondo esas técnicas.

El EEM (o algún otro índice de la exactitud de la medición) permite prevenir un énfasis inadecuado en una sola puntuación numérica. Esta aplicación del EEM es tan importante que cada vez son más las pruebas que expresan las puntuaciones no como un solo número, sino como el margen de puntuaciones dentro del cual es probable que se encuentre la verdadera puntuación del individuo. La Junta Universitaria proporciona datos sobre el EEM y una explicación de su uso no sólo en los materiales distribuidos a consejeros de la educación media y superior, sino también en los reportes de calificación individual que el SAT envía a los examinados. El EEM también se incluye en los materiales que explican a los estudiantes el significado de los resultados que obtuvieron en la prueba. También da información sobre los EEM para interpretar los resultados de los Exámenes de Registro de Graduados (*Graduate Record Examinations*, *Guía* GRE 1995–96).

Interpretación de las diferencias de las puntuaciones. Cuando se evalúan las *diferencias* entre dos puntuaciones es muy importante considerar la confiabilidad de la prueba y los errores de medición. Pensar en términos del intervalo dentro del cual puede fluctuar cada calificación equivale a considerar una supervisión para impedir un acento excesivo de las pequeñas diferencias entre puntuaciones. La precaución resulta deseable lo mismo al comparar las puntuaciones obtenidas por diferentes personas que al equiparar los resultados del mismo individuo en diferentes habilidades. Del mismo modo, es necesario interpretar a la luz de los errores de medición los cambios en las puntuaciones que siguen a la instrucción o a otras variables experimentales.

Una pregunta frecuente acerca de las puntuaciones de las pruebas tiene que ver con la posición relativa del individuo en diferentes áreas. ¿Es mayor la capacidad de Nora para las tareas verbales que para las numéricas?, ¿tiene Tomás más aptitud para las acti-

vidades mecánicas que para las verbales? Si en una batería de aptitudes Nora obtuvo mayor puntuación en la subprueba verbal que en la numérica y Tomás calificó más alto en la subprueba mecánica que en la verbal, ¿qué tan seguros podemos estar de que volverán a hacerlo al ser examinados con otra forma de la batería? En otras palabras, ¿sería posible que las diferencias en las puntuaciones fueran producto de la selección fortuita de los reactivos de las pruebas verbal, numérica y mecánica? Estas preguntas son pertinentes para la interpretación apropiada de los resultados en las baterías de calificaciones múltiples tanto de habilidades como de rasgos de personalidad (Anastasi, 1985a). En los capítulos 8 y 10 (sobre los tests de habilidad) y 13 (acerca de los tests de personalidad) daremos otros ejemplos y un estudio más profundo de los problemas que deben considerarse al interpretar el perfil de puntuaciones en esas baterías.

Debido al interés creciente en la interpretación de los perfiles de calificación, las editoriales elaboraron formas de reporte que permiten la evaluación de las puntuaciones en términos de sus errores de medición. Un ejemplo es la hoja de reporte individual para los Tests de Aptitud Diferencial, que incluye la información ilustrada en la figura 4.6. De este modo, las calificaciones percentilares obtenidas en cada subtest de la batería se grafican como bandas de percentiles alrededor del obtenido. Cada barra percentilar corresponde a una distancia de un EEM a cualquier lado de la puntuación obtenida; por consecuencia, la probabilidad de que la puntuación "verdadera" se encuentre dentro de la barra es aproximadamente de 2 a 1 (.68 a .32). Al interpretar los perfiles se advierte a los usuarios que no den importancia a las diferencias entre las puntuaciones cuyas barras percentilares se sobreponen, en especial si lo hacen en más

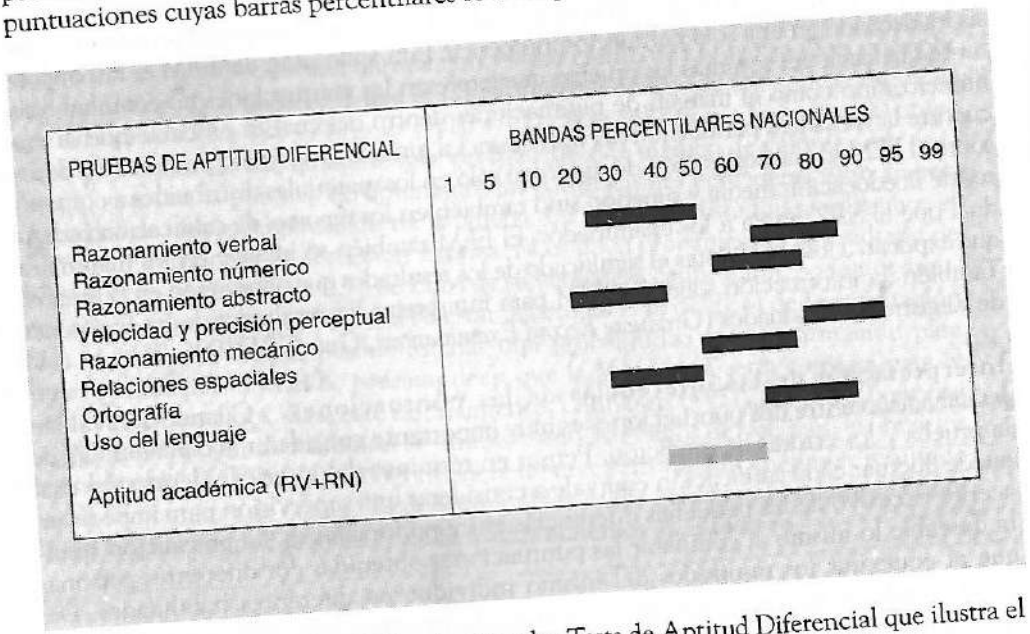


Figura 4.6. Perfil de calificaciones en los Tests de Aptitud Diferencial que ilustra el uso de las bandas percentilares.

(Datos de *Individual Report, Differential Aptitude Tests*, 5ª edición. Copyright © 1990 de The Psychological Corporation. Reproducido con autorización.)

de la mitad de su longitud; por ejemplo, en el perfil de la figura 4.6 es probable que la diferencia entre las puntuaciones de razonamiento verbal y de razonamiento numérico refleje una diferencia genuina en el nivel de habilidad; pero quizá no sea así entre el razonamiento numérico y razonamiento abstracto; mientras que la diferencia entre el razonamiento abstracto y razonamiento mecánico queda en el margen de la duda.

Es bueno recordar que el error estándar de la diferencia entre dos puntuaciones es mayor que el error de medición de cualquiera de esos dos resultados. Esto se deriva del hecho de que los errores aleatorios o fortuitos de *ambas* puntuaciones influyen en esta diferencia. El error estándar de la diferencia entre dos puntuaciones puede encontrarse a partir de los errores estándares de medición de los dos resultados mediante la siguiente fórmula:[9]

$$EE_{dif.} = \sqrt{(EEM_1)^2 + (EEM_2)^2}$$

en la que EE_{dif} es el error estándar de la diferencia entre las dos puntuaciones, y EEM_1 y EEM_2 son los errores estándares de medición de las puntuaciones separadas. Al sustituir el EEM_1 por $DE\sqrt{1 - r_{11}}$ y EEM_2 por $DE\sqrt{1 - r_{22}}$ podemos replantear la fórmula directamente en términos de coeficientes de confiabilidad, de la siguiente manera:

$$EE_{dif.} = DE\sqrt{2 - r_{11} - r_{22}}$$

En esta sustitución, se utilizó la misma *DE* para las pruebas 1 y 2, ya que sus puntuaciones tendrían que expresarse en términos de la misma escala antes de que pudieran compararse.

Podemos ilustrar el procedimiento anterior con los CI verbal y de ejecución de la Escala de Inteligencia para Adultos de Wechsler, Revisada (WAIS-R). La confiabilidad de división por mitades de esos resultados fue de .97 y .93, respectivamente. Los CI de desviación del WAIS-R se expresan en una escala con una media de 100 y una *DE* de 15; por ende, el error estándar de la diferencia entre esas dos puntuaciones puede encontrarse de la siguiente manera:

$$EE_{dif.} = 15\sqrt{2 - .97 - .93} = 4.74$$

Para determinar qué tan grande puede ser una diferencia obtenida por azar a nivel de .05, multiplicamos el error estándar de la diferencia (4.74) por 1.96. El resultado es 9.29, aproximadamente 10 puntos. Así, la diferencia entre el CI Verbal y de Ejecución de un individuo en el WAIS-R debe ser al menos de 10 puntos para ser significativo a nivel de .05.[10]

---

[9] Esta fórmula no debe ser confundida con la fórmula para obtener el error estándar de una diferencia entre dos *medias de grupo*, la cual incluye un término de correlación cuando las dos variables a comparar están correlacionadas. Los *errores de medición* en dos variables son errores aleatorios o fortuitos y por lo tanto se supone que no están correlacionados.

[10] Pueden obtenerse estimaciones más precisas usando las confiabilidades y el EEM que realmente se encuentran en cada grupo de edad. Cuando se calcula de este modo, la mínima diferencia significativa verbal-ejecución al nivel de 0.5, como se informa en el manual de la prueba, va de 8.83 a 12.04; sin embargo, la mayor parte de los valores se encuentran cerca de 10.

APLICACIÓN DE LA CONFIABILIDAD A LAS PRUEBAS DE DESTREZA Y LAS PUNTUACIONES DE CORTE

Recordará que en el capítulo 3 los tests referidos a dominio suelen (pero no necesariamente) evaluar el desempeño en términos de destreza más que en función del grado de rendimiento. Una implicación estadística importante de las pruebas de destreza es una reducción en la variabilidad de las puntuaciones entre personas. Teóricamente, si dos continúan el entrenamiento hasta dominar la habilidad, la variabilidad se reduce a cero. En una sección anterior vimos que a *cualquier* correlación, incluidos los coeficientes de confiabilidad, la afecta la variabilidad del grupo en que se calculó. Conforme disminuye la variabilidad de la muestra también lo hace el coeficiente de correlación. En consecuencia, resultaría inapropiado evaluar la confiabilidad de la mayor parte de los tests referidos a dominio aplicando los procedimientos comunes a un grupo de personas después de que han alcanzado el nivel preestablecido de destreza. En tales condiciones, incluso una prueba muy estable y consistente produciría un coeficiente de confiabilidad cercano a cero.

Esta aparente dificultad para la evaluación de la confiabilidad surge del hecho de no considerar lo que los tests referidos a dominio pretenden medir. En la práctica, los instrumentos se utilizan sobre todo para distinguir entre quienes han adquirido las habilidades y los conocimientos requeridos para determinada actividad y los que no lo han hecho. Al aplicar la prueba puede variar considerablemente, desde obtener la licencia para conducir, ser asignado a una especialidad ocupacional, avanzar a la siguiente unidad en un programa de instrucción individualizada o ser admitido a cierto curso; no obstante, en todas esas situaciones el hecho de que la prueba sea utilizada implica la expectativa de variabilidad individual en el desempeño. Una parte importante de esta variabilidad refleja las diferencias individuales en la cantidad de entrenamiento recibido en las funciones pertinentes.

Hay más de una docena de técnicas distintas para evaluar específicamente la confiabilidad de los tests referidos a dominio (Berk, 1984b; Brennan, 1984; Subkoviak, 1984). Algunas son apropiadas para decisiones simples de dominio-no dominio en las que todos los errores de clasificación se consideran igualmente serios cualquiera que sea su distancia de la puntuación de corte. En estos casos, es posible aplicar test y retest con formas paralelas para encontrar el porcentaje de personas para las cuales se tomó la misma decisión en ambas ocasiones. Estos datos se analizan luego, al calcular los índices de acuerdo con valores significativos. Otros procedimientos toman en consideración las puntuaciones reales obtenidas en ambas ocasiones y proporcionan indicadores que reflejan la desviación de cada persona por encima o por debajo de cualquier puntuación de corte. La elección de un procedimiento especial debe tener en cuenta la naturaleza y los usos de la prueba. La bibliografía especializada ha examinado ampliamente consideraciones adecuadas (véase Berk, 1984a; Feldt y Brennan, 1989).

Validez: Conceptos básicos

La validez de los instrumentos de medición tiene que ver con lo *que* miden y con *qué tan bien* lo hacen; nos indican qué se puede inferir a partir de sus resultados. A este respecto, debe tenerse el cuidado de no aceptar el nombre de la prueba o el test como indicador de lo que mide, pues aunque son títulos cortos, convenientes para propósitos de identificación, casi todos son demasiado generales y vagos para indicar el área de conducta que cubren; no obstante, cada vez son mayores los esfuerzos por utilizar nombres más específicos y que puedan definirse de manera empírica. El rasgo medido por determinada prueba sólo puede definirse mediante el examen de las fuentes objetivas de información y las operaciones empíricas utilizadas para establecer su validez. Más aún, la validez no puede expresarse en términos generales, no es posible decir en abstracto que es "alta" o "baja", sino que hay que referirse al uso particular para el que se planea utilizar el instrumento.

En principio, todos los procedimientos utilizados para determinar la validez se interesan en las relaciones entre la ejecución en las pruebas y otros factores observados independientemente de las características de la conducta considerada. Para investigar estas relaciones se han empleado distintos métodos, cuyos nombres tradicionalmente se han concentrado en aspectos de la validez que son importantes para distintos usos de la prueba. En la medida en que se han desarrollado y expandido las aplicaciones de los instrumentos, también se han modificado los conceptos de validez (Anastasi, 1986a; Messick, 1988, 1989).

EVOLUCIÓN DE LOS CONCEPTOS DE VALIDEZ DE LAS PRUEBAS

Uno de los primeros usos de las pruebas fue la evaluación de lo que los individuos habían aprendido en determinadas áreas de contenido. En la actualidad, se aplican a los exámenes escolares de fin de cursos y las pruebas presentadas para obtener la licencia de manejo o para desempeñar ciertas ocupaciones. En general, para evaluar esta categoría de pruebas se compara su contenido con el del área que pretenden probar. La aproximación descriptiva sigue siendo importante en la validación de los instrumentos para algunas aplicaciones, la veremos en otra sección de este capítulo.

Cuando la examinación pasó a su segunda etapa, el énfasis cambió a la predicción, ¿cómo responderán diferentes personas a una situación dada, ahora o en algún momento futuro?, ¿cómo reaccionará este individuo en diferentes situaciones específicas? Se designó como criterio al desempeño en la situación para la cual se pretende predecir la conducta. En este caso, la validez del instrumento generalmente se informa como el coeficiente de correlación entre los resultados del test y una medida de criterio directa e independiente. Este procedimiento es en especial apropiado para el uso de las pruebas en la selección o colocación de individuos en programas educativos, empleos o programas particulares de tratamiento. De este modo, para una prueba de aptitud mecánica el criterio puede ser el desempeño laboral como maquinista; para una de aptitud académica, las calificaciones escolares, y para una prueba de neuroticismo, las valoraciones de los compañeros o alguna otra información disponible sobre el comportamiento del individuo en diversas situaciones.

La corriente actual en la historia de la evaluación refleja dos tendencias principales: (1) un fortalecimiento de la orientación teórica, y (2) una estrecha vinculación entre la teoría y la verificación psicológicas mediante la comprobación empírica y experimental de las hipótesis. Estas tendencias son tan evidentes en la elaboración y la validación de los instrumentos como en el conjunto de las otras áreas de la psicología (Anastasi, 1992a, 1992b, 1995). Un resultado de estas corrientes es el reconocimiento creciente del valor de los constructos para describir y comprender la conducta humana. Los constructos son categorías amplias que se derivan de los rasgos comunes que comparten las variables conductuales observables directamente; pero se trata de entidades teóricas que por su parte no pueden ser observadas de manera directa. El interés en los constructos llevó a la introducción de lo que al principio se consideró la tercera categoría de validez, a saber, la validez de constructo (AERA, APA, NCME, 1985; APA, AERA, NCME, 1974; Cronbach y Meehl, 1955). A la larga y en la medida en que especifica lo que la prueba mide, la validez de constructo llegó a reconocerse como la validez fundamental e incluyente de concepto. Los procedimientos de validación predictiva y de contenido se encuentran entre las muchas fuentes de información que contribuyen a la definición y la comprensión de los constructos evaluados por la prueba. Al mismo tiempo, proporcionan información que es valiosa por derecho propio y destacan en la evaluación de las pruebas para determinados usos. De ahí que los conceptos (y términos) hayan sobrevivido a pesar de su integración en el concepto más amplio de validez de constructo.

PROCEDIMIENTOS DE LA DESCRIPCIÓN DEL CONTENIDO

Naturaleza. Los procedimientos de validación por la descripción del contenido comprenden principalmente el examen sistemático del contenido de la prueba para

determinar si cubre una muestra representativa del área de conducta que debe medirse. Esta forma de validación se utiliza sobre todo en los instrumentos diseñados para medir qué tan bien ha dominado el individuo una habilidad o un curso de estudio.

Puede parecer que basta con la simple inspección del contenido de la prueba para establecer su validez para dicho propósito; por ejemplo, una prueba de multiplicación, ortografía o contabilidad parecería válida por definición si está constituida por reactivos de multiplicación, ortografía o contabilidad. Pero la solución no es tan sencilla. Un problema es el muestreo adecuado del universo de reactivos. El área de conducta por examinar debe analizarse sistemáticamente para garantizar que los reactivos cubran todos los aspectos importantes y en la proporción correcta. Es fácil cargar en exceso las pruebas con los aspectos del campo que más se prestan para la preparación de reactivos objetivos. Por ello, resulta conveniente describir de antemano toda el área considerada en lugar de hacerlo después de que la prueba está lista; por ejemplo, un examen educativo bien formulado debe cubrir los objetivos de la instrucción y no sólo su temario. En consecuencia, hay que definir el contenido de manera amplia para que, además del conocimiento real, incluya objetivos importantes como la aplicación de principios y la interpretación de datos. Más aún, la validez depende más de la relevancia que las respuestas del individuo tengan para el área conductual considerada que la importancia aparente del contenido del reactivo. La simple inspección de la prueba no siempre basta para revelar los procesos empleados por los examinados al presentarla.

En relación con el área muestreada por la prueba, es importante prevenir cualquier tendencia a la sobregeneralización; por ejemplo, una prueba de ortografía compuesta por reactivos de opción múltiple puede medir la habilidad para reconocer las palabras escritas correcta e incorrectamente, pero no puede suponerse que también mida la habilidad para tomar un dictado, la frecuencia de los errores ortográficos en un trabajo de redacción y otros aspectos de la habilidad ortográfica (Ahlström, 1964; Knoell y Harris, 1952). Otro problema procede de la posible inclusión de factores irrelevantes en los resultados; así, la habilidad para comprender instrucciones verbales o la rapidez para ejecutar tareas rutinarias sencillas pueden afectar indebidamente una prueba diseñada para medir el aprovechamiento en matemáticas o mecánica.

Procedimientos específicos. La validez de contenido se introduce desde el inicio en la prueba mediante la elección de reactivos apropiados. Para las pruebas educacionales, la preparación de los reactivos es precedida por una revisión cuidadosa y sistemática de textos y resúmenes importantes para el curso y por la consulta con expertos en la materia. Sobre la base de la información recabada se establecen las *especificaciones de la prueba* que deben seguir los redactores de los reactivos y en las que tienen que precisar las áreas o temas del contenido, los objetivos o procesos educativos que han de probarse y la importancia relativa de temas y procesos individuales. Las especificaciones finales deben indicar el número de reactivos de cada clase que hay que preparar para cada tema; por ejemplo, la evaluación de la habilidad de lectura puede incluir la comprensión del vocabulario en el contexto, la comprensión literal del contenido y la extracción de inferencias correctas de la información proporcionada. También puede muestrear material de diversas fuentes, como ensayos, poemas, artículos periodísticos o instructivos para manejar equipo. Una prueba de matemáticas puede cubrir habilidades

de cálculo, la resolución de problemas presentados verbalmente y la aplicación de los procesos aprendidos a contextos nuevos.

La reseña de la validación del contenido en el manual de una prueba educativa de aprovechamiento debe incluir la descripción de los procedimientos seguidos para asegurar que el contenido del instrumento es apropiado y representativo. Si en la preparación de éste participaron expertos en la materia, debe informarse de su número y calificación profesional. Si fungieron como jueces en la clasificación de reactivos, es necesario referir las instrucciones que recibieron así como el grado de acuerdo entre ellos. Como los planes de estudio y los contenidos de los cursos cambian con el tiempo, es particularmente deseable proporcionar las fechas en las que se consultó a los expertos. También tiene que proporcionarse información relacionada con el número y la naturaleza de resúmenes y textos consultados, incluidas las fechas de publicación.

Para complementar la validación de contenido de una prueba de aprovechamiento académico suele seguirse una serie de procedimientos empíricos. Pueden supervisarse los resultados totales y el desempeño en reactivos particulares para calificar el progreso. En general, se retienen los reactivos que muestran las mayores ganancias en los porcentajes de niños que avanzan de los grados inferiores a los superiores. Cuando resulta apropiado, se utilizan otros procedimientos complementarios, como el análisis de las categorías de errores más comunes en la prueba y la observación de los métodos de trabajo empleados por los examinados. Para esto último, es posible pedir a los estudiantes que "piensen en voz alta" mientras resuelven cada problema. Para verificar la contribución de la rapidez se anota el número de quienes no lograron terminar la prueba o se aplica alguno de los métodos más refinados que vimos en el capítulo 4. Para detectar la posible influencia irrelevante de la habilidad para leer instrucciones en la ejecución de la prueba, las calificaciones obtenidas en el examen se correlacionan con una prueba de comprensión de lectura. Por otro lado, si la prueba fue diseñada para medir la comprensión de lectura, entregar el cuestionario sin el pasaje de lectura en el que se basó mostrará cuántas preguntas puede contestar el examinado a partir de la información que ya poseía o con la ayuda de otras señales irrelevantes (Scherich y Hanna, 1977).

Aplicaciones. En especial, cuando es apoyada por supervisiones empíricas como las ilustradas, la validación de contenido proporciona una técnica apropiada de evaluación, pues nos permite responder a dos preguntas que son básicas para la validez de las pruebas de rendimiento académico y ocupacional: (1) ¿cubre la prueba una muestra representativa de las habilidades y los conocimientos especificados?, (2) ¿el desempeño en la prueba está razonablemente libre de la influencia de variables irrelevantes? La validación de contenido es apropiada sobre todo para las pruebas referidas al dominio que estudiamos en el capítulo 3. Como el desempeño en estos instrumentos se interpreta en términos del significado del contenido, es evidente que la validación de contenido es un requisito primordial para su uso efectivo; no obstante, otras formas de evidencia de validez también son importantes para una evaluación completa de su efectividad (véase Hambleton, 1984b).

La validación de contenido también se aplica a ciertas pruebas ocupacionales (que se revisan en el capítulo 17) diseñadas para la selección y clasificación de personal. Estas evidencias de validación son adecuadas cuando la prueba es una muestra del empleo real o de otro que requiera las mismas habilidades y conocimiento que demanda

el puesto. En esos casos, debe realizarse un análisis exhaustivo del puesto para demostrar una similitud estrecha entre las actividades del puesto y la prueba. Schoenfeldt, Schoenfeldt, Acker y Perlson (1976) ofrecen una clara y detallada explicación de la aplicación de esos procedimientos de validación al desarrollo de una prueba de lectura industrial. Trabajando junto a supervisores y empleados, los investigadores analizaron el nivel de lectura requerido para ingresar a una compañía manufacturera en términos del tema y las habilidades de comprensión, y luego formularon reactivos que correspondieran a dichos requisitos. Esta aproximación suele emplearse al elaborar pruebas para los empleados del gobierno estadounidense a nivel federal y estatal (Hardt, Eyde, Primoff y Tordy, 1981; Menne, McCarthy y Menne, 1976; Primoff y Eyde, 1988; Tordy, Eyde, Primoff y Hardt, 1976).

Por otro lado, en el caso de los tests de aptitud y de personalidad, la validación de contenido no sólo suele ser inapropiada, sino incluso engañosa. Aunque es obvio que en las etapas iniciales de la elaboración de cualquier instrumento deben hacerse consideraciones sobre la relevancia y representatividad del contenido, la validación final de los tests de aptitud o personalidad debe verificarse empíricamente mediante los procedimientos que describiremos en las siguientes secciones. En comparación con las pruebas de aprovechamiento, estos tests tienen menor parecido intrínseco con el dominio de conducta que pretenden muestrear. En consecuencia, su contenido hace más que revelar las hipótesis que llevaron a quien las realizó a elegir ciertos contenidos para medir un rasgo específico. Las hipótesis tienen que ser confirmadas empíricamente para establecer la validez de los instrumentos.

A diferencia de las pruebas de aprovechamiento, los tests de aptitud y de personalidad no se basan en un curso específico de educación o en un conjunto uniforme de experiencias de las cuales pueda extraerse el contenido del instrumento; por consecuencia, es probable que en estas pruebas se encuentren más diferencias en los métodos de trabajo o en los procesos psicológicos empleados por los individuos al responder a los mismos reactivos. Ello supone que una misma prueba puede medir funciones diferentes en distintas personas, lo que hace prácticamente imposible determinar las funciones psicológicas medidas por la prueba a partir de la inspección de su contenido; por ejemplo, los estudiantes de posgrado pueden resolver un problema en términos verbales o matemáticos, mientras que un mecánico llegará a la misma solución en términos de visualización espacial; o una prueba que en estudiantes de secundaria mide el razonamiento aritmético, cuando se aplica a universitarios puede medir apenas las diferencias individuales en la velocidad de los cálculos.

Validez de facie. La validez de contenido no debe confundirse con la validez de facie (conocida también como validez aparente). Esta última no es validez en el sentido técnico, es decir, no se refiere a lo que la prueba verdaderamente mide, sino a lo que parece medir. La validez de facie alude a si la prueba "parece válida" a los examinados que la presentan, al personal administrativo que decide sobre su uso y a otros observadores sin capacitación técnica. En esencia, la cuestión de la validez de facie tiene que ver con el *rapport* y las relaciones públicas. Aunque el uso común del término "validez" a este respecto puede resultar confuso, la validez de facie es en sí misma un rasgo deseable de los instrumentos; así, cuando las pruebas originalmente diseñadas para niños y formuladas dentro de las aulas empezaron a extender su uso a los adul-

tos, fue común que enfrentaran críticas y resistencia por carecer de validez de facie. Es indudable que si el contenido de la prueba parece irrelevante, inapropiado, tonto o infantil, el resultado sea una pobre cooperación, cualquiera que sea su verdadera validez. En particular en la evaluación de adultos, para que una prueba funcione no basta con que sea objetivamente válida, también debe parecerlo. La validez de facie también influye en la aceptación de la prueba en decisiones legislativas y judiciales, así como en el juicio que de ella haga el público en general.

En un innovador programa de investigación sobre el punto de vista del examinado (citado en el capítulo 1), Baruch Nevo y sus asociados incluyeron la investigación de la validez de facie (B. Nevo, 1985, 1992; B. Nevo y Sfez, 1985). Al principio les llamó la atención observar que, a pesar de su probable contribución a las actitudes prevaleciente hacia las pruebas, eran muy pocas las investigaciones sobre la validez de facie. Propusieron luego una evaluación cuantitativa de la validez aparente al hacer que examinados y otros interesados sin conocimiento psicométrico calificaran la conveniencia de cierta prueba para determinado uso; los mismos procedimientos pueden emplearse para calificar reactivos de una sola prueba o una batería. A este respecto resultan ilustrativos los datos obtenidos al analizar las respuestas de 1 385 estudiantes israelíes a un cuestionario de retroalimentación de examinados aplicado después de un examen de admisión a la universidad compuesto por seis pruebas. Los resultados mostraron un prometedor acuerdo entre calificadores, confiabilidad en el retest y diferenciación entre pruebas y entre subgrupos de aspirantes que planeaban especializarse en distintas áreas. Se recomendó que los manuales incluyeran regularmente datos cualitativos y cuantitativos sobre la validez de facie.

La validez de facie o aparente a menudo puede mejorarse mediante el simple recurso de replantear los reactivos de modo que parezcan relevantes y plausibles en el medio particular en el que serán usados; por ejemplo, si se construye una prueba simple de razonamiento aritmético para aplicarla a maquinistas, los reactivos deben plantearse en términos de operaciones con las máquinas y no de "cuántas naranjas pueden comprarse con X pesos" u otros problemas tradicionales de los textos escolares. De igual modo, es posible elaborar una prueba de aritmética para personal naval en la terminología náutica sin alterar con ello las funciones medidas. Por supuesto, la validez de facie no debe considerarse como sustituto de la validez determinada objetivamente. No puede suponerse que al mejorar la validez de facie de una prueba mejore su validez objetiva, pero tampoco se puede presumir que cuando se modifica una prueba para aumentar su validez de facie su validez objetiva quede inalterada. La validez de la prueba en su forma final siempre debe verificarse de manera directa.

PROCEDIMIENTOS DE CRITERIO-PREDICCIÓN

Validación concurrente y predictiva. Los procedimientos de validación de criterio-predicción indican la efectividad de la prueba para predecir el desempeño del individuo en actividades específicas. La medida de criterio contra la que se validan los resultados del instrumento puede obtenerse aproximadamente al mismo tiempo que los resultados de la prueba o después de un intervalo establecido. Los estándares de examinación ("*Testing Standards*") de 1985 utilizan estas relaciones temporales entre el

criterio y la prueba como base para diferenciar entre la validación concurrente y la predictiva. El término "predicción" se utiliza en sentido amplio para referirse a la suposición que puede hacer la prueba sobre cualquier situación de criterio, o bien en el sentido más restringido de anticipación sobre un intervalo. La expresión "validez predictiva" se emplea en este último sentido, y la información que proporciona es muy pertinente para las pruebas usadas en la selección y clasificación de personal. La contratación de personal, la selección de estudiantes para su ingreso a la universidad o a escuelas profesionales y la asignación de personal militar a programas de capacitación ocupacional son algunos ejemplos de las decisiones que requieren el conocimiento de la validez predictiva de los instrumentos. Otros ejemplos incluyen el uso de las pruebas para descartar a los individuos que probablemente desarrollen trastornos emocionales en ambientes de tensión o para identificar a los pacientes psiquiátricos que es más probable que se beneficien de cierta terapia.

Cuando no es factible extender los procedimientos de validación por el tiempo requerido para obtener la validación predictiva o una muestra adecuada de preselección para los propósitos del examen, se emplea la validación concurrente como sustituto. En estos casos, la prueba se aplica a un grupo del que ya se tienen los datos de criterio; por ejemplo, se comparan los resultados obtenidos en la prueba por universitarios con su promedio académico en el momento del examen, o los resultados obtenidos por empleados con su éxito actual en el trabajo.

Por otro lado, para ciertos usos de los tests la validación concurrente resulta más conveniente y puede justificarse por derecho propio. La distinción lógica entre la validación predictiva y concurrente no se basa en el tiempo, sino en los objetivos del examen. La validación concurrente es adecuada para las pruebas que se emplean para *diagnosticar* el estado actual más que para predecir los resultados futuros. La diferencia puede ilustrarse al preguntar "¿califica Ortega como un buen piloto?" o "¿posee Ortega los requisitos para convertirse en un buen piloto?" La primera pregunta requiere de validación concurrente; la segunda, de validación predictiva.

Como el criterio para la validación concurrente siempre está disponible en el momento de aplicar el examen, podría cuestionarse qué función cumple la prueba en esas situaciones. Básicamente, los instrumentos proporcionan un sustituto más simple, rápido y menos costoso para los datos del criterio; por ejemplo, si el criterio consiste en la observación continua de un paciente durante dos semanas de hospitalización, una prueba que pueda distinguir a los casos normales de los casos perturbados o dudosos reduciría notablemente el número de personas que requieren tan prolongada observación.

Contaminación del criterio. Una precaución esencial al encontrar la validez de una prueba consiste en asegurar que sus propios resultados no influyan en la condición del criterio de cualquier individuo; por ejemplo, si un maestro o el supervisor de una planta industrial saben que determinado individuo obtuvo un pésimo resultado en un test de aptitud, ese conocimiento puede influir en la calificación que otorguen al estudiante o en la clasificación que asignen al trabajador; en contraste, quien obtuvo una puntuación elevada puede recibir el beneficio de la duda cuando se preparan sus calificaciones académicas o clasificaciones laborales. Es evidente que estas influencias elevarán la correlación entre los resultados obtenidos en la prueba y el criterio de una manera por completo espuria o artificial.

Esta posible fuente de error en la validación de la prueba se conoce como contaminación del criterio, ya que las calificaciones del criterio quedan "contaminadas" por el conocimiento que el calificador tiene de los resultados del instrumento. Para evitar este error es absolutamente esencial que ninguna de las personas que participan en la asignación de las calificaciones del criterio tenga conocimiento alguno de los resultados obtenidos por los examinados en la prueba. Es por ello que los resultados que se emplean para evaluarla deben mantenerse estrictamente confidenciales. En ocasiones resulta difícil convencer a los maestros, empleadores, oficiales del ejército, etc., de que esa precaución es esencial. En su prisa por utilizar toda la información disponible para tomar decisiones prácticas, pueden no percatarse de que los resultados deben dejarse a un lado hasta que los datos del criterio maduren y pueda supervisarse la validez.

Medidas de criterio. Una prueba puede validarse si se la compara con tantos criterios como usos específicos existan para ella. Cualquier método utilizado para evaluar la conducta en cualquier situación puede brindar una medida de criterio para algún propósito particular; sin embargo, los criterios empleados para encontrar las distintas clases de validez que se informan en los manuales caen en unas cuantas categorías comunes. Entre los criterios más empleados para validar los tests de inteligencia se encuentra cierto índice de *aprovechamiento académico*, razón por la cual a menudo se describen de manera más precisa como medidas de aptitud académica. Los índices específicos utilizados como medidas de criterio comprenden las calificaciones escolares, los resultados de pruebas de aprovechamiento, los registros de promoción y de graduación, los honores y reconocimientos especiales y las valoraciones de maestros o instructores sobre la "inteligencia" de los individuos. En tanto que es muy probable que las calificaciones obtenidas en un medio escolar se distorsionen por la ejecución académica del individuo, pueden clasificarse más apropiadamente con el criterio de aprovechamiento académico.

Los diversos índices de aprovechamiento académico han proporcionado datos del criterio en todos los niveles educativos, desde la educación básica hasta la superior. Aunque se emplean principalmente en la validación de los tests de inteligencia, también se han usado como criterio para ciertos tests de habilidad múltiple y de personalidad; por ejemplo, en la validación de cualquiera de las pruebas de ingreso a la universidad, un criterio común es el promedio obtenido durante el primer año de estudios. Esta medida es la calificación promedio obtenida en todos los cursos durante el primer año; cada calificación es sopesada por el número de puntos del curso por quienes fue recibida.

En el caso de los adultos que no asisten a la escuela, suelen emplearse los años de escolaridad completa como una variante del criterio de aprovechamiento académico. En general, se espera que los individuos más inteligentes prolonguen su educación por más tiempo y que los otros abandonen antes la escuela. Esta suposición fundamenta el criterio de que la escala educativa funge como una influencia progresivamente selectiva, que elimina a los que no son capaces de continuar más allá de algún peldaño. Aunque indudablemente es cierto que los graduados universitarios, por ejemplo, representan un grupo más selecto que los graduados de la educación básica, la relación entre la escolaridad y la aptitud académica está lejos de ser perfecta. Especialmente en los niveles educativos superiores, las variables económicas, sociales, motivacionales y otros factores no educativos pueden influir en la continuación de la educación del individuo. Más aún, con la validación concurrente es difícil desentrañar

las relaciones causales. ¿En qué medida las diferencias en los resultados de los tests de inteligencia se deben a las diferencias en la escolaridad? ¿En qué medida predice la prueba las diferencias individuales en el progreso educativo subsecuente? Estas preguntas se responden sólo cuando el instrumento se aplica antes de que los datos del criterio hayan madurado, como en la validación predictiva.

En el desarrollo de las pruebas de aptitudes especiales, un criterio frecuente se basa en el *desempeño en la capacitación especializada*; por ejemplo, las pruebas de aptitud mecánica pueden validarse si se compara sus resultados con el rendimiento final en un curso del oficio. Algunos cursos de escuelas comerciales, tales como mecanografía o contabilidad, proporcionan criterios para los tests de aptitud en esas áreas. De igual modo, se ha empleado el desempeño en las escuelas de arte o de música para validar las pruebas de aptitud musical o artística. Varias pruebas de aptitud profesional han sido validadas en términos del aprovechamiento en las escuelas de leyes, medicina, odontología, ingeniería y otras. En el caso de las pruebas elaboradas para usarse en un programa específico de examinación, los registros de la capacitación son una fuente frecuente de datos del criterio. Un magnífico ejemplo es la validación de las pruebas de selección de pilotos para la fuerza aérea estadounidense, que compara sus resultados con el desempeño en los vuelos básicos de entrenamiento. La ejecución en los programas de capacitación también suele utilizarse como criterio para la validación de las pruebas en otras especialidades ocupacionales militares y en algunos estudios de validación industrial.

Entre los índices específicos del desempeño en la capacitación que se utilizan como criterio pueden mencionarse las pruebas de aprovechamiento administradas al terminar el curso, las calificaciones formalmente asignadas, las valoraciones de los instructores y la culminación exitosa en contraste con la eliminación del programa. Para determinar la validez de las baterías de aptitudes múltiples como predictores diferenciales, es frecuente que se comparen sus resultados con las calificaciones obtenidas en secundarias o cursos universitarios específicos; por ejemplo, los resultados en una prueba de comprensión verbal pueden compararse con las calificaciones obtenidas en los cursos de redacción, los resultados de visualización espacial con las calificaciones en geometría, etcétera.

En lo que respecta al uso general de los registros de capacitación como medidas del criterio, resulta útil distinguir entre los criterios intermedios y los finales. Así, en el desarrollo de una prueba de selección de pilotos de la fuerza aérea estadounidense o de una prueba de aptitud médica, los criterios últimos serían el desempeño en combate y el éxito en la práctica médica. Por supuesto, se requeriría mucho tiempo para la maduración de esos datos, por lo que es dudoso que en la práctica pueda obtenerse un criterio último verdadero, y, aunque fuera posible estaría afectado por tantos factores no controlados que su utilidad sería nula; por ejemplo, sería difícil evaluar el grado relativo de éxito de médicos de diferentes especialidades que ejercen en diversas partes del país. Es por eso que suelen emplearse criterios intermedios como medidas, como los registros del desempeño en alguna etapa de la capacitación.

Para muchos propósitos, la medida de criterio más satisfactoria es la que se basa en registros de seguimiento del *desempeño laboral* real, un criterio que hasta cierto punto se ha empleado en la validación de los tests de inteligencia general y de personalidad y en mayor medida en la validación de pruebas de aptitudes especiales. También es un criterio común en la validación de instrumentos diseñados para trabajos específicos. Los "trabajos" en cuestión varían tanto en nivel como en categoría, incluido el trabajo en los negocios, la industria, las profesiones y las fuerzas armadas. Aunque es probable que casi

todas las medidas del desempeño en el trabajo no representen los criterios últimos, sí proporcionan un buen criterio intermedio para muchos propósitos de examinación, lo que los hace preferibles a los registros de capacitación. Por otro lado, la medición del desempeño en el empleo no permite tanta uniformidad en las condiciones como es posible en la capacitación. Más aún, como por lo general requiere un seguimiento más prolongado, es probable que el criterio del desempeño en el trabajo ocasione una disminución en el número de los participantes disponibles. Debido a la variación en la naturaleza de empleos nominalmente similares en distintas organizaciones, los manuales que informan los datos de validez contra el criterio del puesto no sólo deben describir las medidas de criterio que emplearon, sino también las tareas realizadas por los trabajadores.

La validación por el método de *grupos contrastados* generalmente incluye un criterio compuesto que refleja las influencias selectivas acumuladas y no controladas de la vida cotidiana. Este criterio se basa finalmente en la supervivencia dentro de un grupo particular en comparación con la eliminación de éste; por ejemplo, para comprobar la validez de una prueba de aptitud musical o mecánica se comparan los resultados obtenidos por los estudiantes inscritos en una escuela de música o de ingeniería, respectivamente, con los resultados de estudiantes no admitidos en la secundaria o la universidad. Por supuesto, para seleccionar a los grupos contrastados puede usarse cualquier criterio (como calificaciones escolares, valoraciones, desempeño laboral), y se eligen los extremos de la distribución de las medidas elegidas; sin embargo, los grupos contrastados que se incluyen en esta categoría se han ido diferenciando gradualmente por la operación de las múltiples exigencias de la vida cotidiana. El criterio a consideración es, por ende, más complejo y su definición resulta menos clara que los que estudiamos arriba.

El uso del método de grupos contrastados es muy común en la validación de los tests de personalidad; por ejemplo, al validar una prueba de rasgos sociales puede compararse el desempeño de vendedores o ejecutivos con el de oficinistas o ingenieros. La suposición que fundamenta al procedimiento es que, en relación con muchos rasgos sociales, los individuos que han ingresado y permanecido en ocupaciones como las ventas o el trabajo ejecutivo superarán como grupo a los que se encuentran en campos como el trabajo de oficina o la ingeniería. De modo similar, los estudiantes universitarios que han participado en muchas actividades extracurriculares pueden compararse con los que no han participado en ninguna durante un periodo comparable de asistencia a la universidad. Los grupos ocupacionales han sido muy usados en el desarrollo y la validación de las pruebas de interés, como el Formulario de Intereses Vocacionales de Strong (*Strong Vocational Interest Blank*, SVIB), así como en la preparación de escalas de actitud. En la validación de escalas de actitud se han empleado, entre otros, grupos políticos, religiosos, geográficos y otros que generalmente son conocidos por ostentar puntos de vista distintos sobre ciertos temas.

En la validación empírica de las pruebas referidas a dominio, además de los procedimientos comunes de validación de contenido, se han utilizado varias adaptaciones del método de grupos contrastados (Hambleton, 1984b). Con este propósito se compara para el desempeño en la prueba de grupos que difieren en el nivel de educación relevante. Si se emplea una calificación de habilidad puede hacerse un análisis 2×2, que compara la proporción de resultados de éxito y fracaso antes y después de la instrucción (Panell y Laabs, 1979). Pueden hacerse comparaciones similares si la prueba se aplica a escolares de un grado inferior y uno superior al grado en el que se enseña el concepto o la habilidad particular que evalúa la prueba. Si se dispone de los resultados

después de diferentes periodos de educación, se encontrará la correlación entre el desempeño real y la cantidad de instrucción.

En el desarrollo de ciertos tests de personalidad, el *diagnóstico psiquiátrico* se utiliza como base para la selección de reactivos y como evidencia de la validez del instrumento. El diagnóstico psiquiátrico puede ser un criterio satisfactorio siempre que se funde en la observación prolongada y la historia de caso detallada más que en una entrevista o un examen psiquiátrico superficial. En el último de los casos, no hay razón para esperar que el diagnóstico psiquiátrico sea superior al resultado de la prueba como indicador de la condición emocional del individuo. No debe considerarse al diagnóstico psiquiátrico como medida de criterio, sino como un indicador o predictor cuya propia validez tiene que ser determinada.

Ya hablamos, en relación con otras categorías de criterios, de las *valoraciones* de los maestros, instructores de cursos especializados y supervisores laborales. A éstas pueden agregarse las opiniones de oficiales en situaciones militares, las calificaciones que los consejeros escolares dan a los estudiantes y las valoraciones de los compañeros de trabajo, de la escuela, del club, etc. Las apreciaciones o calificaciones que hemos visto representan simplemente una técnica complementaria para obtener información respecto a criterios como el rendimiento académico, el desempeño en la capacitación especializada o el éxito en el trabajo. Ahora, se está considerando su uso como el núcleo de la medida de criterio; en tales circunstancias, son las propias valoraciones o calificaciones las que definen el criterio. Más aún, dichas calificaciones no se limitan a la evaluación de un aprovechamiento específico, sino que incluyen el juicio personal de un observador en relación con cualquier rasgo que las pruebas psicológicas pretenden medir. Así, los participantes en la muestra de validación pueden ser calificados en características como dominancia, ingenio mecánico, originalidad, liderazgo u honestidad.

Se han empleado las valoraciones en la validación de casi cualquier clase de prueba, pero su utilidad es mayor en la obtención de criterios para los tests de personalidad, ya que en esta área es mucho más difícil encontrar criterios objetivos. Esto es cierto sobre todo en el caso de rasgos distintivamente sociales en los que las calificaciones basadas en el contacto personal pueden formar el criterio más lógico. Aunque las valoraciones pueden estar sujetas a muchos errores de juicio, cuando se obtienen en condiciones cuidadosamente controladas representan una fuente valiosa de datos del criterio. En el capítulo 16 se consideran las técnicas para mejorar la precisión de las valoraciones y para reducir las formas más comunes de error.

Por último, es común que se citen las correlaciones entre la nueva prueba y *pruebas validadas previamente disponibles* como evidencia de validez. Cuando el nuevo instrumento es una forma abreviada o simplificada de una prueba ya existente, esta última puede considerarse como la medida de criterio. De esta manera, una prueba de lápiz y papel puede validarse contrastándola con una prueba de aprovechamiento más elaborada y cuya validez ya haya sido establecida, o bien una prueba de grupo se valida comparándola con una prueba individual; por ejemplo, en repetidas ocasiones se ha empleado el Stanford-Binet como criterio en la validación de tests de grupo. En tales condiciones, el nuevo instrumento puede considerarse, en el mejor de los casos, como una aproximación gruesa del anterior. Debe observarse que a menos que la nueva prueba represente un sustituto más sencillo o más breve de la prueba anterior, es inapropiado el uso de la última como criterio.

Un avance importante en la elaboración de tests en las décadas de los ochenta y los noventa ha centrado su atención en el *análisis del criterio*, un aspecto tradicionalmente descuidado en la investigación de las pruebas. Con los años, se han elevado voces dispersas para señalar la necesidad de realizar una investigación sistemática de los criterios, pero la puesta en práctica ha dado pocos frutos (L. R. James, 1973; Tenopyr, 1986). Incluso en los proyectos bien diseñados que emplearon el análisis cuidadoso del puesto como guía para el desarrollo del instrumento, los resultados tuvieron poco efecto sobre la medida de criterio empleada en la validación subsecuente de la prueba. Por lo general, se aceptaba que el criterio "estaba ahí", y con demasiada frecuencia se representaba con un índice global del desempeño en el trabajo basado en calificaciones o registros de producción total. Ahora se reconoce ampliamente que la validez de una prueba se explora de mejor manera si primero se identifican los constructos más importantes en la realización del trabajo y luego se eligen o elaboran instrumentos cuyos resultados los evalúen (J. P. Campbell, 1990a; J. P. Campbell, McHenry y Wise, 1990; L. V. Jones y Appelbaum, 1989; Messick, 1995). El Proyecto de Selección y Clasificación del Ejército Estadounidense, conocido como "Proyecto A" (J. P. Campbell, 1990b), es un buen ejemplo de la investigación minuciosa del criterio como primer paso en el desarrollo de una batería de pruebas. Debido a su importancia general para el uso industrial y organizacional de las pruebas, en el capítulo 17, nos ocuparemos más a fondo de este proyecto a gran escala de siete años.

Generalización de la validez. A menudo se utiliza la validez de criterio-predicción en la validación local de estudios que pretenden evaluar la efectividad de una prueba para un programa específico. Es lo que se hace, por ejemplo, cuando determinada compañía desea evaluar una prueba para la selección de personal para uno de sus puestos o cuando a una universidad le interesa determinar qué tan bien puede predecir cierta prueba de aptitud académica el desempeño de sus estudiantes. La validez predictiva se caracteriza mejor como la validez práctica del instrumento para un propósito específico.

Cuando en los estudios de validación industrial se empezó a correlacionar las pruebas estandarizadas de aptitud con el desempeño en los puestos supuestamente similares, se encontró una gran variabilidad en los coeficientes de validez (Ghiselli, 1959, 1966). Al emplear como criterio las calificaciones obtenidas en diversos cursos escolares se observó una variabilidad similar entre los coeficientes de validez (G. K. Bennett, Seashore y Wesman, 1984). Esos descubrimientos generaron un enorme pesimismo en torno a la posibilidad de generalizar la validez de la prueba a situaciones distintas. Hasta mediados de los setenta, la "especificidad situacional" de los requisitos psicológicos era considerada como una seria limitación de la utilidad de las pruebas estandarizadas para la selección de personal; sin embargo, en un complejo análisis estadístico del problema, Schmidt, Hunter y sus colegas demostraron que gran parte de la varianza entre los coeficientes de validez puede ser un ardid o engaño estadístico que resulta del pequeño tamaño de la muestra, la poca confiabilidad del criterio y la restricción del rango en las muestras de empleados.[1]

[1] Este trabajo formó parte de un programa continuo de investigación informado en muchos artículos y monografías. Las publicaciones más importantes para el presente tema incluyen (sin limitarse) Pearlman, Schmidt y Hunter (1980), Schmidt, Gast-Rosenberg y Hunter (1980), Schmidt y Hunter (1977), Schmidt, Hunter y Pearlman (1981) y Schmidt, Hunter, Pearlman y Shane (1979).

Las muestras industriales disponibles para la validación de las pruebas suelen ser demasiado pequeñas para producir una estimación estable de la correlación entre predictor y el criterio. Por la misma razón, los coeficientes obtenidos pueden ser demasiado pequeños para alcanzar significación estadística en la muestra empleada, por lo que no logran aportar evidencias sobre la validez del instrumento. Se ha estimado que alrededor de la mitad de las muestras de validación usadas en estudios industriales no incluye más de 40 o 50 casos (Schmidt, Hunter y Urry, 1976). Con muestras tan pequeñas, técnicamente no es posible aplicar la validación de criterio-predicción.

Al aplicar sus técnicas de reciente desarrollo a los datos de muchas muestras extraídas de un gran número de especialidades ocupacionales, Schmidt, Hunter y sus colaboradores pudieron demostrar que la validez de las pruebas de aptitud verbal, numérica y de razonamiento puede generalizarse entre ocupaciones mucho más de lo que se había reconocido. La varianza de los coeficientes de validez que se encontraban en los primeros estudios industriales demostró no ser mayor a lo que se habría esperado por azar, incluso cuando las funciones del puesto en particular parecían ser muy distintas entre los puestos. Evidentemente, el desempeño exitoso de una amplia variedad de tareas ocupacionales depende en un grado sustancial de un núcleo común de habilidades cognoscitivas. Las pruebas incluidas en esos estudios cubrían principalmente la clase de contenido y habilidades muestreadas en las pruebas tradicionales de inteligencia y de aptitud académica. Parecería que este grupo de conocimientos y habilidades cognoscitivas es un buen predictor del desempeño en las actividades académicas y ocupacionales requeridas en las sociedades tecnológicamente avanzadas; no obstante, en general pueden tomarse decisiones de selección más precisas al considerar los resultados obtenidos en dos o tres grupos cognoscitivos amplios, de preferencia complementados con las medidas de las habilidades técnicas específicas para trabajos particulares (Hartigan y Wigdor, 1989; L. L. Wise, McHenry y Campbell, 1990; Zeidner y Johnson, 1991).

Meta-análisis. Los procedimientos estadísticos empleados para investigar la generalización de la validez permiten integrar los descubrimientos de diferentes estudios. Estos procedimientos hacen posible combinar los datos de investigaciones realizadas en momentos o en lugares diferentes, o de informaciones publicadas en distintos estudios. Tales procedimientos, inicialmente designados como meta-análisis, empezaron a usarse en la investigación psicológica durante los setenta (Glass, 1976; Schmidt y Hunter, 1977), aunque en otras ciencias han estado en uso durante varias décadas (Hartigan y Wigdor, 1989, capítulo 6). En la psicología, el meta-análisis ha recibido cada vez mayor atención como sustituto de la revisión tradicional de la bibliografía (Lipsey y Wilson, 1993; Schmidt, 1992). Por lo común, estas revisiones señalaban el número de estudios que encontraron efectos estadísticamente significativos en, por ejemplo, las diferencias entre las medias de los grupos experimental y de control o las correlaciones entre los resultados de la prueba y otras variables. Con este procedimiento, a menudo se perdían resultados prometedores porque las muestras empleadas en los estudios eran demasiado pequeñas para arrojar diferencias significativas.

Al combinar los descubrimientos publicados de varios estudios y sopesarlos hasta donde fuera posible sobre la base de las características sustantivas y metodológicas relevantes de cada estudio, el meta-análisis puede revelar descubrimientos positivos sus-

tanciales. Otra ventaja es que permite calcular los *efectos del tamaño, la magnitud o la medida*. Por razones tanto teóricas como prácticas, la magnitud estimada de una diferencia o una correlación es más útil que la simple demostración de que es significativamente mayor que cero.

Las décadas de los ochenta y los noventa han presenciado el rápido resurgimiento de la investigación meta-analítica casi en cualquier campo de la psicología. Sus aplicaciones a la investigación de la selección y clasificación de personal son las que han atraído la mayor atención (véase el capítulo 17). El interés en el meta-análisis crece a paso firme y de continuo se establecen procedimientos más refinados. Aunque persiste cierta controversia acerca de las técnicas, los resultados principales no varían apreciablemente entre procedimientos.[2]

PROCEDIMIENTOS DE IDENTIFICACIÓN DEL CONSTRUCTO

El término "validez de constructo" fue oficialmente introducido en el léxico psicométrico en 1954 en las *Recomendaciones técnicas para las pruebas psicológicas y las técnicas de diagnóstico* (APA, 1954), que constituyeron la primera edición de los actuales *Estándares de examinación*. Al año siguiente apareció la primera exposición detallada de la validez de constructo en un artículo de Cronbach y Meehl (1955). Las controversias sobre la validez de constructo que siguieron (y que continúan hoy con el mismo vigor) han servido para hacer más explícitas las implicaciones de sus procedimientos y para proporcionar una lógica sistemática de su uso. La validación de constructo ha centrado la atención en la función que cumple la teoría psicológica en la elaboración de la prueba y en la necesidad de formular hipótesis que puedan ser comprobadas o refutadas en el proceso de validación. También ha estimulado la búsqueda de nuevas formas de obtener los datos sobre la validez. Aunque algunas de las técnicas empleadas en la investigación de la validez de constructo se conocen desde hace mucho tiempo, el campo de aplicación se ha ampliado para admitir una mayor variedad de procedimientos.

La validez de constructo de un instrumento es el grado en el que puede afirmarse que mide un constructo o rasgo teórico. La aptitud académica, la comprensión mecánica, la fluidez verbal, la rapidez de la marcha, el neuroticismo y la ansiedad son algunos ejemplos de dichos constructos. La validación de constructo requiere de la acumulación gradual de diversas fuentes de información; cada constructo se deriva de las interrelaciones establecidas entre medidas conductuales y se forma para organizar y dar cuenta de las concordancias observadas en la respuesta. Cualquier dato que arroje luz sobre la naturaleza del dato considerado y sobre las condiciones que afectan su desarrollo y manifestaciones constituye una evidencia apropiada para su validación. En las siguientes secciones veremos ejemplos concretos de técnicas que contribuyen a la identificación del constructo.

---

[2] Aplicaciones recientes, explicaciones detalladas de los procedimientos y evaluaciones críticas pueden encontrarse en Hartigan y Wigdor (1989), Hedges (1988), Hunter y Schmidt (1990), L. R. James, Demaree, Mulaik y Ladd (1992), L. V. Jones y Appelbaum (1989), R. Rosenthal (1991), Schmidt (1992), Schmidt *et al.* (1993) y Schmidt, Ones y Hunter (1992). Para una introducción sencilla a los procedimientos estadísticos, véase F. M. Wolf (1986). Para una visión más amplia del uso del meta-análisis en la investigación conductual, véase Cook *et al.* (1992), Cooper y Hedges (1994), Hasselblad y Hedges (1995) y Wachter y Straf (1990).

Cambios en el desarrollo. La *diferenciación por edad* es un importante criterio utilizado en la validación de una serie de tests tradicionales de inteligencia. Instrumentos como el Stanford-Binet y la mayor parte de las pruebas para preescolares se contrastan con la edad cronológica para determinar si las puntuaciones muestran un incremento progresivo conforme aumenta la edad. Como se espera que, durante la niñez, las habilidades aumenten con la edad, se argumenta que si la prueba es válida sus resultados deberían mostrar dicho incremento. El mismo concepto de una escala cronológica de inteligencia, como la iniciada por Binet, se basa en la suposición de que la "inteligencia" aumenta con la edad, al menos hasta la madurez.

Desde luego, el criterio de diferenciación por edad no puede aplicarse a las funciones que no muestren cambios claros y consistentes con el paso del tiempo; por ejemplo, se ha descubierto que su uso es limitado en el área de la medición de la personalidad. Más aún, debe notarse que, incluso cuando es aplicable, la diferenciación por edad es una condición necesaria pero no suficiente para la validez. Luego pues, si los resultados de la prueba no mejoran con la edad, ese descubrimiento quizá indique que el instrumento no es una medida válida de las habilidades que debería muestrear. Por otro lado, demostrar que una prueba mide algo que se incrementa con la edad no define con precisión el área cubierta por la prueba. Una medida de estatura o peso también debería mostrar incrementos regulares con la edad, aunque sobra decir que no se le consideraría en un test de inteligencia.

Subrayemos un último punto relativo a la interpretación del criterio de edad. Un test psicológico validado contra dicho criterio mide características conductuales que se incrementan con la edad en las condiciones existentes en el entorno en el que se estandarizó el instrumento. Como diferentes culturas pueden estimular y fomentar el desarrollo de características conductuales disímiles, no puede suponerse que el criterio de diferenciación por edad sea universal. Como cualquier otro criterio, está circunscrito por el contexto cultural particular del que fue derivado.

Los análisis del desarrollo también son básicos para la validación de constructo de las escalas ordinales piagetianas que tratamos en los capítulos 3 y 9. Una suposición fundamental de dichas escalas es el *patrón secuenciado* de desarrollo, en el que la consecución de las primeras etapas en el desarrollo del concepto es un requisito para la adquisición de las habilidades conceptuales posteriores. Existe así una jerarquía intrínseca en el contenido de estas escalas, por lo que la validación de constructo de las escalas ordinales debe incluir datos empíricos de la estabilidad de la secuencia de etapas sucesivas, lo que supone supervisar la ejecución de los niños en diferentes niveles en el desarrollo de cualquier concepto examinado, como la conservación o la permanencia del objeto. ¿Los niños que muestran dominio del concepto en cualquier nivel también exhiben dominio en los niveles inferiores?

Correlaciones con otros tests. En ocasiones se citan las correlaciones entre una nueva prueba y otros instrumentos similares como evidencia de que la nueva prueba mide aproximadamente la misma área general de conducta que otras que llevan el mismo nombre, como "tests de inteligencia" o "pruebas de aptitud mecánica". A diferencia de las correlaciones encontradas en la validez de criterio-predicción, estas correlaciones deben ser moderadamente altas, pero no demasiado. Si la nueva prueba se correlaciona

muy bien con cualquier otra ya disponible, sin agregar ventajas como la brevedad o facilidad de aplicación, entonces el nuevo instrumento es una duplicación innecesaria.

Otro uso de las correlaciones con otras pruebas consiste en demostrar que la nueva está relativamente libre de la influencia de ciertos factores irrelevantes; por ejemplo, un test de aptitud especial o uno de personalidad no deberían tener una correlación elevada con tests de inteligencia general o de aptitud académica. De modo similar, la comprensión de lectura no debería afectar apreciablemente el desempeño en dichos instrumentos. De acuerdo con ello, las correlaciones con los tests de inteligencia general, lectura o comprensión verbal en ocasiones se aportan como evidencia indirecta o negativa de validez. En esos casos, las correlaciones elevadas pondrían a la prueba bajo sospecha, aunque, por otro lado, las correlaciones bajas no son garantía de validez. Advierta que este uso de las correlaciones con otros instrumentos es similar al de las técnicas complementarias que vimos al estudiar los procedimientos de descripción de contenido.

Análisis factorial. Desarrollado como un medio para identificar rasgos psicológicos, el análisis factorial es particularmente relevante para los procedimientos de validación de constructo. En esencia, se trata de una refinada técnica estadística para analizar las interrelaciones de los datos conductuales; por ejemplo, si se aplican 20 pruebas a 300 personas, el primer paso consiste en calcular las correlaciones de cada instrumento con el resto. Una inspección de la tabla de las 190 correlaciones resultantes puede revelar ciertas agrupaciones entre las pruebas, lo que indica la localización de rasgos comunes. Así, si pruebas como las de vocabulario, analogías, opuestos y completamiento de oraciones muestran altas correlaciones entre sí y correlaciones bajas con las otras pruebas, podemos inferir tentativamente la presencia de un factor de comprensión verbal. Como el análisis mediante la inspección de una tabla de correlación es difícil e incierto, se han elaborado técnicas estadísticas más precisas para localizar los factores comunes que se requieren para explicar las correlaciones obtenidas, y que estudiaremos en el capítulo 11 en relación con su uso en la investigación sobre la naturaleza de la inteligencia, de la que se originaron.

En el proceso del análisis factorial, se reduce el número de variables o categorías en cuyos términos puede describirse el desempeño de cada individuo a un número relativamente pequeño de factores o rasgos comunes. En el ejemplo citado, cinco o seis factores pueden ser suficientes para explicar las correlaciones entre las 20 pruebas. En lugar de que cada individuo sea descrito en términos de los 20 resultados originales, la descripción se hace de acuerdo con los resultados que haya obtenido en cinco o seis factores. Un propósito importante del análisis factorial es simplificar la descripción de la conducta, reduciendo el número de categorías a unos cuantos factores o rasgos comunes.

Una vez que los factores han sido identificados, sirven para describir la composición factorial de las pruebas. Cada instrumento puede entonces caracterizarse en función de los factores principales que determinan sus calificaciones, junto con el peso o carga de cada uno y la correlación de la prueba con cada factor, que suele expresarse como la *validez factorial* de la prueba. De esta manera, si el factor de comprensión verbal correlaciona .66 con una prueba de vocabulario, la validez factorial de esta prueba de vocabulario como medida del rasgo de comprensión verbal es .66. Observe que la validez factorial es en esencia la correlación de la prueba con lo que exista en común en un

grupo de pruebas u otros índices de conducta. Por supuesto, el conjunto de variables analizado puede incluir datos de la prueba y ajenos a ella. Es posible aprovechar las valoraciones y otras medidas de criterio, junto con otras pruebas, para explorar la composición factorial de alguna prueba y para definir los rasgos comunes que mide.

Consistencia interna. En la descripción de algunos tests, especialmente en el campo de la personalidad, se afirma que han sido validados por el método de consistencia interna. La característica esencial de este método es que el criterio no es otro que la calificación total del propio instrumento. En ocasiones se utiliza una adaptación del método de grupos contrastados, seleccionando grupos extremos sobre la base de la calificación total en la prueba. En cada reactivo se compara la ejecución del grupo de criterio superior con el desempeño del grupo inferior. Los reactivos que no logran mostrar una proporción significativamente mayor de "aciertos" (o respuestas esperadas) en el grupo superior que en el inferior se consideran inválidos y se eliminan o revisan. También se emplean con este propósito los procedimientos de correlación; por ejemplo, puede calcularse la correlación biserial entre el resultado total de la prueba y el "éxito o fracaso" en cada reactivo y elegir únicamente los reactivos que producen correlaciones significativas entre reactivo y prueba (item-test). Se dice que una prueba cuyos reactivos fueron seleccionados por este método tiene consistencia interna, ya que al igual que la prueba, cada reactivo distingue a los que responden en la misma dirección.

Otra aplicación del criterio de consistencia interna incluye la correlación de las calificaciones de los subtests con el resultado total; por ejemplo, muchos tests de inteligencia constan de subpruebas que se aplican por separado (como las de vocabulario, aritmética, completación de figuras, etc.) y cuyos resultados se combinan para encontrar el resultado total. En la elaboración de esas pruebas, a menudo se correlacionan las calificaciones de cada subtest con la calificación total y se elimina cualquier subtest cuya correlación con éste sea demasiado baja. Las correlaciones de los subtests restantes se muestran luego como evidencia de la consistencia interna del instrumento.

Es evidente que las correlaciones de consistencia interna, se basen en reactivos o en subtests, son esencialmente medidas de homogeneidad. El grado de homogeneidad de la prueba tiene cierta relevancia para su validez de constructo porque contribuye a caracterizar al área de conducta o rasgo que muestrea; no obstante, la contribución de los datos de consistencia interna a la validación de la prueba es limitada. Si se carece de datos externos, es poco lo que puede saberse acerca de lo que mide.

Validez convergente y discriminante. En un minucioso análisis de la validación de constructo, D. T. Campbell (1960) señaló que para demostrar la validez de constructo, no basta con demostrar que una prueba tiene una correlación elevada con otras variables con las que en teoría debe hacerlo, sino también que *no* tiene una correlación significativa con variables de las que se supone debe diferir. En un artículo anterior, D. T. Campbell y Fiske (1959) llamaron al primer proceso validación convergente y al último validación discriminante. La correlación de una prueba de razonamiento cuantitativo con las calificaciones obtenidas luego en un curso de matemáticas sería un ejemplo de validación convergente. Para la misma prueba, la validez discriminante se haría evidente con una correlación baja e insignificante con los resultados de una

prueba de comprensión de lectura, ya que la habilidad de lectura es una variable irrelevante en una prueba diseñada para medir razonamiento cuantitativo.

Recuerde que estudiamos el requisito de una baja correlación con variables irrelevantes al tratar los procesos complementarios y preventivos seguidos en la validación de contenido. La validación discriminante es especialmente pertinente en el caso de las pruebas de personalidad en las que las variables irrelevantes pueden afectar de muchas maneras los resultados.

Campbell y Fiske (1959) propusieron un diseño experimental sistemático para la aproximación dual de validación convergente y discriminante que llamaron la *matriz de multimétodo-multirrasgo*. En esencia, el procedimiento requiere de la evaluación de dos o más rasgos mediante dos o más métodos. Un ejemplo hipotético proporcionado por Campbell y Fiske permitirá ilustrar el procedimiento. La tabla 5.1 muestra todas las correlaciones posibles entre los resultados obtenidos cuando tres rasgos se miden con tres métodos. Los rasgos pueden representar tres características de personalidad como (A) dominancia, (B) sociabilidad y (C) motivación de logro. Los tres métodos pueden ser (1) un inventario autodescriptivo, (2) una técnica proyectiva y (3) las opiniones de los compañeros o pares. Así, A_1 indica las puntuaciones de dominancia obtenidas en el inventario de autorreporte, A_2 los resultados de dominancia en la técnica proyectiva y C_3 representa las calificaciones dadas por los compañeros sobre la motivación de logro.

Las correlaciones hipotéticas mostradas en la tabla 5.1 incluyen coeficientes de confiabilidad (entre paréntesis sobre la diagonal principal) y coeficientes de validez (en negritas sobre las tres diagonales cortas). En estos coeficientes de validez, se correlacionan las calificaciones obtenidas para el mismo rasgo mediante diferentes métodos; cada medida se compara así con otras medidas independientes del mismo rasgo, como en el procedimiento de validación ya conocido. La tabla comprende también las correlaciones entre rasgos *diferentes* medidos por el *mismo* método (en los triángulos continuos) y las correlaciones entre rasgos *diferentes* medidos por métodos *distintos* (en los triángulos punteados). Para que la validez de constructo sea satisfactoria, los coeficientes de validez deben ser obviamente mayores a las correlaciones entre rasgos diferentes medidos por métodos distintos, y también deberían ser mayores que las correlaciones encontradas entre rasgos distintos medidos por el mismo método; por ejemplo, la correlación entre los resultados de dominancia obtenidos con un inventario autodescriptivo y los obtenidos con una técnica proyectiva deberían ser mayores que la correlación entre las calificaciones de dominancia y sociabilidad obtenidas en el inventario. Si esta última correlación, que representa un método común de varianza, fuera elevada, indicaría, por ejemplo, que algún factor común irrelevante, como la habilidad para entender las preguntas o el deseo de presentarse bajo una luz favorable en todos los rasgos, influyó indebidamente en los resultados que el sujeto obtuvo en el inventario.

Intervenciones experimentales. Los experimentos sobre el efecto de variables seleccionadas en los resultados de la prueba constituyen otra fuente de datos para la validación de constructo; por ejemplo, una estrategia para comprobar la validez de una prueba que se pretende emplear en un programa de educación personalizada consistiría en comparar las calificaciones obtenidas en el pretest y el postest. La lógica de este método supone que en el pretest (previo a la educación pertinente) deben obtenerse bajas puntuaciones y que en el postest los resultados deben ser elevados. Esta re-

Tabla 5.1

Matriz hipotética multimétodo-multirrasgo

		Método 1			Método 2			Método 3		
	Rasgos	A_1	B_1	C_1	A_2	B_2	C_2	A_3	B_3	C_3
	A_1	(.89)								
Método 1	B_1	.51	(.89)							
	C_1	.38	.37	(.76)						
	A_2	**.57**	.22	.09	(.93)					
Método 2	B_2	.22	**.57**	.10	.68	(.94)				
	C_2	.11	.11	**.46**	.59	.58	(.84)			
	A_3	**.56**	.22	.11	**.67**	.42	.33	(.94)		
Método 3	B_3	.23	**.58**	.12	.43	**.66**	.34	.67	(.92)	
	C_3	.11	.11	**.45**	.34	.32	**.58**	.58	.60	(.85)

Nota. Las letras A, B, C se refieren a los rasgos; los subíndices 1, 2, 3 a los métodos. Los coeficientes de validez (un solo rasgo, métodos diferentes) son los tres conjuntos de diagonales de números en negritas; los coeficientes de confiabilidad (mismo rasgo, mismo método) son los números entre paréntesis sobre la diagonal principal. Los triángulos continuos encierran las correlaciones entre rasgos múltiples y un solo método; los triángulos punteados encierran las correlaciones entre rasgos y métodos múltiples.

(Tomado de Campbell y Fiske, 1959, pág. 82. Copyright 1959 por la American Psychological Association. Reproducido con autorización).

lación también puede revisarse para los reactivos individuales de la prueba. De manera ideal, la mayoría de los examinados debería fallar un reactivo en el pretest y aprobarlo en el postest; los reactivos que se fallan en ambos son demasiado difíciles, y los que se aciertan en los dos son demasiado sencillos para los propósitos del examen. Si una proporción considerable de examinados acierta un reactivo en el pretest y lo falla en el postest, obviamente hay algo mal en el reactivo, la educación o en ambos.

Por ejemplo, podemos aplicar a un grupo de sujetos una prueba diseñada para medir la predisposición a la ansiedad y someterlos luego a una situación destinada a ponerlos ansiosos, como tener que presentar un examen en condiciones distractoras y estresantes. Después correlacionamos los resultados obtenidos en la prueba con índices fisiológicos y con otras expresiones de ansiedad durante y después del examen. O bien, podemos evaluar una hipótesis diferente sobre la prueba de ansiedad aplicándola antes y después de una experiencia que genere ansiedad para luego ver si las calificaciones se elevan significativamente en la segunda aplicación de la prueba. En tal expe-

rimento, un resultado positivo indicaría que las calificaciones de la prueba reflejan el nivel actual de ansiedad. De manera similar, pueden diseñarse experimentos para comprobar otras hipótesis concernientes al rasgo medido por una determinada prueba.

Modelamiento de ecuaciones estructurales. Además de permitir la identificación de los constructos en los resultados de las pruebas y en el criterio de ejecución, un importante avance en la validación de las pruebas fue la consideración de las relaciones entre constructos y la vía por la que éstos afectan la ejecución (J. P. Campbell, 1990a; Messick, 1989; Schmidt, Hunter y Outerbridge, 1986); por ejemplo, el interés de una persona en un área en particular puede afectar su desempeño laboral al aumentar su aprendizaje en el conocimiento real relevante, al facilitarle la adquisición de las habilidades de procedimiento requeridas o al desarrollar la motivación necesaria para ejercer y sostener un esfuerzo máximo al llevar a cabo las actividades requeridas. Investigar *cómo* es que un constructo o un rasgo personal identificado conduce a un buen o mal desempeño contribuye sustancialmente a la comprensión de por qué una prueba tiene una elevada o baja validez en una situación dada. Facilita dicho análisis un procedimiento estadístico conocido como modelamiento de ecuaciones estructurales, cuyo uso ha aumentado de manera impresionante en los ochenta y los noventa. Véase, por ejemplo, el establecimiento de la revista *Structural Equation Modeling* (1994). Este procedimiento está muy relacionado con algunas versiones del análisis de trayectorias, y ambos suelen conocerse informalmente como "modelamiento causal".[3]

¿Cómo se originó y con qué propósito se diseñó el modelamiento de ecuaciones estructurales? En la estadística elemental, pronto se aprende que correlación no indica causalidad. Un ejemplo familiar es la correlación espuria con la edad; por ejemplo, entre los escolares de seis a 14 años es probable que la correlación entre la estatura y el aprovechamiento en aritmética sea elevada, pero eso no permite concluir que una variable afecta a la otra. La correlación resulta principalmente de la edad, que a su vez está asociada con el monto de escolaridad. En un esfuerzo por investigar las relaciones causales, en las décadas de los sesenta y los setenta los investigadores comenzaron a usar diseños experimentales de series de tiempos (D. T. Campbell y Stanley, 1966; Cook y Campbell, 1976, pp. 284–293). Así, para analizar la influencia recíproca causal entre las actitudes de una estudiante hacia las matemáticas y su desempeño en la materia, se recaban medidas de las actitudes y el desempeño en matemáticas en dos momentos. Luego, se obtiene la correlación de las series de tiempos entre la actitud hacia las matemáticas en el momento 1 y el desempeño en la materia en el momento 2, y entre el desempeño en matemáticas en el momento 1 y la actitud hacia la materia en el momento 2. El tamaño de ambas correlaciones indicaría la fuerza relativa de la influencia en cada dirección. Durante algunos años, este método parecía una forma prometedora de evaluar el efecto recíproco de dos variables.

Sin embargo, al poco tiempo los análisis lógicos y estadísticos revelaron serias debilidades en el uso de las correlaciones en las series de tiempos. Aunque el diseño básico era

[3] Para evitar las implicaciones filosóficas acerca de la causa última o completa de cualquier suceso, los psicólogos han preferido expresiones más neutrales como afirmar que A determina, influye o afecta a B. No obstante, en ocasiones se emplea el término "causal" para referirse a todas esas relaciones, con la suposición de que se comprenden sus restricciones (véase, por ejemplo, L. R. James, Mulaik y Brett, 1982, capítulo 1; P. A. White, 1990).

excelente, es probable que el uso de correlaciones simples de rango cero distorsionara los resultados y llevara a conclusiones incorrectas acerca de relaciones causales (Rogosa, 1980). Entre las fuentes de error del procedimiento está el hecho de no tomar en consideración, primero, las correlaciones entre las variables iniciales y entre las variables subsecuentes; segundo, la confiabilidad de las variables y su estabilidad en el tiempo, y tercero, la posible contribución de variables no medidas, como la edad y la escolaridad en el ejemplo citado. El modelamiento de ecuaciones estructurales proporciona los medios para evitar estas dificultades, ya que utiliza ecuaciones de regresión para predecir las variables dependientes a partir de las independientes en los diseños de series de tiempos u otros modelos causales. El procedimiento utiliza correlaciones parciales para encontrar los coeficientes de regresión, lo que le permite incorporar todas las correlaciones entre las variables; asimismo, considera los errores de medición y de muestreo e incluye las previsiones para reconocer al menos la posibilidad de otras variables causales no medidas (Bentler, 1988; L. R. James *et. al.*, 1982; Loehlin, 1992; Rogosa, 1979).

En el modelamiento de ecuaciones estructurales el primer paso consiste en diseñar un modelo de las relaciones causales hipotéticas que quieren probarse. Es importante que el modelo se base en una profunda familiaridad con los conocimientos acerca de las variables y la situación investigadas, por lo que las relaciones hipotetizadas deben resultar teóricamente racionales. La comprobación del modelo se realiza al resolver un conjunto de ecuaciones[4] simultáneas de regresión lineal. En el modelamiento causal suele haber más ecuaciones que incógnitas, lo que permite encontrar soluciones para varios modelos alternativos. Cada modelo se compara con la matriz original de correlación empírica para supervisar la conveniencia del ajuste, y es posible encontrar un ajuste casi igualmente bueno en varios modelos causales (MacCallum, Wegener, Uchino y Fabrigar, 1993). Esos modelos estadísticamente equivalentes representan diferentes trayectorias causales y, por ende, brindan explicaciones alternativas de los efectos observados de manera empírica. El investigador, sobre la base del conocimiento que tenga de la situación, evaluará la credibilidad y el significado sustantivo de los modelos alternativos.

Otra característica del modelamiento de ecuaciones estructurales es que suele calcular relaciones causales entre constructos más que entre variables aisladas; por ejemplo, para evaluar la actitud de un estudiante hacia las matemáticas pueden usarse varios indicadores, como las medidas de interés, de orientación a la meta, el autoconcepto sobre la aptitud para las matemáticas y otras variables afectivas pertinentes. La varianza común entre esos indicadores definiría entonces un constructo de la actitud del individuo hacia las matemáticas, que luego se relacionaría con el desempeño subsecuente en la materia. El uso de constructos proporciona estimaciones más estables y confiables que cancelan los errores y las varianzas específicas de los indicadores separados.

En la actualidad, son varias las aproximaciones metodológicas al modelamiento de ecuaciones estructurales, así como las modificaciones y los refinamientos del procedimiento considerado (véase por ejemplo, Anderson y Gerbing, 1988; Bentler, 1990; Bollen y Long, 1993; Breckler, 1990; Cole, Maxwell, Arvey y Salas, 1993; James, 1980;

[4] Para una introducción detallada al procedimiento, véase Bollen (1989) y Loehlin (1992). Los cálculos reales pueden ser realizados con programas de cómputo disponibles como LISREL (Hayduk, 1988; Jöreskog y Sörbom, 1986, 1989) y EQS (Bentler, 1985).

Mulaik *et al.*, 1989). No obstante que aún se encuentra en proceso de desarrollo, el modelamiento de ecuaciones estructurales es un procedimiento prometedor para la combinación de las aproximaciones teórica, experimental y estadística. Ya ha sido aplicado a problemas en áreas como la psicología del desarrollo, de la personalidad, social, industrial y educativa (por ejemplo, Graves y Powell, 1988; L. A. James y L. R. James, 1989; MacCallum y Browne, 1993; McCardle, 1989; Parkerson, Lomax, Schiller y Walberg, 1984; Shavelson y Bolus, 1982). También están en camino esfuerzos para la unificación y simplificación del procedimiento (por ejemplo, Jöreskog y Sörbom, 1993).

Contribuciones de la psicología cognoscitiva. La década de los setenta presenció un acercamiento entre la psicología experimental y la psicometría que así empieza a hacer aportaciones importantes a la comprensión de los constructos evaluados por los tests de inteligencia y otras aptitudes muy definidas (Ronning, Glover, Conoley y Witt, 1987; R. E. Snow y Lohman, 1989). Ya en los cincuenta los psicólogos cognoscitivos empezaron a aplicar los conceptos del procesamiento de información al estudio de la solución de problemas en el ser humano. Algunos investigadores elaboraron programas de cómputo que realizaran esos procesos y simularan el pensamiento humano, programas que repiten el desempeño de personas con diferentes niveles de habilidad, lo que hizo posible predecir el número y la clase de errores cometidos y el tiempo requerido para diferentes respuestas. Al diseñar un programa, el investigador suele empezar con un análisis de la tarea, el cual puede consistir en la introspección sobre el método empleado para resolver el problema, hacer que los sujetos piensen en voz alta o utilizar procedimientos de observación más sofisticados. Al comparar la ejecución de la computadora con la de niños y adultos o bien con la de expertos y novatos mientras resuelven el mismo problema, los investigadores someten a prueba su hipótesis de lo que hace la gente cuando realiza las tareas. Entre las tareas investigadas con esos métodos se incluyen rompecabezas convencionales, problemas de lógica, juegos de ajedrez, problemas de álgebra, de física y de diagnóstico en la práctica médica (Chi, Glaser y Farr, 1988; J. H. Larkin, McDermott, Simon y Simon, 1980a, 1980b; Newell y Simon, 1972; Simon, 1976).

Las variables identificadas por estas investigaciones abarcan procesos (habilidades de procedimiento) y conocimiento declarativo (hechos e información). Los modelos cognoscitivos especifican los procesos intelectuales empleados para realizar la tarea, la forma de organización de los procesos, el almacenamiento del conocimiento relevante y la forma en que se representa en la memoria y se recupera cuando se necesita. También se está concediendo gran atención a lo que ha sido llamado proceso ejecutivo o metacognición, que se refiere al control que el individuo ejerce sobre su elección de procesos, representaciones y estrategias para realizar la tarea. En los setenta, algunos psicólogos cognoscitivos empezaron a aplicar ese análisis de tareas y técnicas de simulación por computadora a la exploración de lo que miden las pruebas de inteligencia. Los investigadores se aproximaron al problema desde distintos ángulos (véase Resnick, 1976; Sternberg, 1981, 1984, 1985b), y la investigación ha ido contribuyendo al avance en la elaboración y el uso de las pruebas.

Las implicaciones de la investigación cognoscitiva para la validación de constructo resultan especialmente claras en el trabajo de Embretson (1983, 1986, 1995a). Al observar que la validación tradicional de constructo ha tenido un alcance demasiado

restringido, Embretson propuso dos aspectos principales en la validación de las pruebas: (1) la representación del constructo, y (2) el espacio nomotético. Por lo común, la validación del constructo se ha concentrado con el espacio nomotético, que se refiere a las relaciones del desempeño en la prueba dentro de una "red nomotética" de variables. Por lo general, esas relaciones se investigan mediante las correlaciones de los resultados de la prueba con otras medidas, que pueden incluir el criterio de ejecución y otros datos de la vida real.

Por otro lado, la representación del constructo tiene por objeto identificar los componentes específicos del procesamiento de información y los almacenes de conocimiento requeridos para realizar las tareas establecidas por los reactivos de la prueba. Para conducir dicho análisis puede realizarse experimentalmente una *descomposición de la tarea*.[5] Los ejemplos de posibles procedimientos comprenden la manipulación de la complejidad de la tarea, la presentación de tareas parciales o la presentación de señales que alteren los requerimientos de la tarea. Se han formulado modelos matemáticos para medir la contribución de diferentes componentes de respuesta a la ejecución en el reactivo. Otro procedimiento de gran uso para el análisis de las tareas cognoscitivas es el *análisis del protocolo* (Ericsson, 1987; Ericsson y Simon, 1993; van Someren, Barnard y Sandberg, 1994), que pide a los individuos que "piensen en voz alta" mientras realizan una tarea o resuelven un problema. Las tareas pueden consistir en hacer una multiplicación mental, recordar detalles de un acontecimiento pasado o localizar la causa del mal funcionamiento de un equipo al responder a reactivos sucesivos en un test de habilidad. Un coproducto de este procedimiento es el posible descubrimiento de que el mismo reactivo puede evocar procesos cognoscitivos diferentes en examinados con experiencia y antecedentes distintos.

¿Qué podemos concluir de la contribución que la psicología cognoscitiva ha hecho hasta ahora a la validación de constructo? Aunque aún se encuentra en una etapa exploratoria, las aproximaciones del procesamiento de información han aportado conceptos heurísticos para guiar la investigación posterior y han llevado la atención a los *procesos de respuesta* en contraste con la tendencia tradicional de la investigación psicométrica a concentrarse en los productos finales del pensamiento. El análisis de la ejecución en la prueba en términos de procesos cognoscitivos específicos sin duda deberá mejorar nuestra comprensión de lo que miden las pruebas. Más aún, analizar el desempeño individual a nivel de los procesos elementales a la larga permitirá la identificación precisa de las fuentes de debilidad y fortaleza de cada individuo y, por consecuencia, aumentará el uso diagnóstico de las pruebas (Embretson, 1987, 1994; Estes, 1974; Pellegrino y Glaser, 1979; Sternberg y Weil, 1980), lo que a su vez facilitará la confección de programas de entrenamiento para las necesidades individuales. En resumen, la relación entre las aproximaciones psicométrica y cognoscitiva puede caracterizarse, primero, como *complementaria* desde el punto de vista de la investigación y práctica aplicada; en este caso, cada una es la aproximación adecuada para diferentes propósitos de evaluación. Segundo, desde el punto de vista de la teoría y la investigación básica, la relación puede caracterizarse como *recíproca*; cada una ayuda a aclarar y enriquecer a la otra, y, en combinación, ambas aumentan nuestra comprensión de la conducta inteligente.

[5] Para más detalles, véase Butterfield, Nielsen, Tangen y Richardson (1985), Embretson (1985b) y Sternberg (1977, 1980).

RECAPITULACIÓN E INTEGRACIÓN

Comparación de los procedimientos de validación. Hemos considerado distintas formas de preguntar qué tan válida es una prueba. Para identificar con claridad los rasgos distintivos de los diferentes procedimientos de validación, hagamos el ejercicio de aplicarlos, uno a la vez, a una prueba que conste de 50 problemas aritméticos. En la tabla 5.2 se ilustran cuatro formas de aplicar esta prueba junto con el procedimiento de validación adecuado para cada caso. Este ejemplo pone de relieve el hecho de que la elección del procedimiento de validación depende del uso que pretenda hacerse de los resultados. Así, si se quiere usar una prueba de aprovechamiento para predecir el desempeño subsecuente en un nivel educativo superior, como es el caso de los procedimientos de admisión a la universidad, debe ser evaluada comparando su resultado con el criterio del rendimiento académico posterior más que en términos de su validez de contenido.

Inclusividad de la validación de constructo. Los ejemplos de la tabla 5.2 se concentran en las diferencias entre los distintos procedimientos de validación; sin embargo, una consideración más detenida de esos procedimientos muestra que la validez de contenido, de criterio y de constructo no corresponden a categorías distintas o lógicamente coordinadas. Por el contrario, la validez de constructo es un concepto inclusivo que abarca a los otros. Todas las técnicas específicas para el análisis del contenido y para medir relaciones con el criterio, que estudiamos en las primeras secciones del capítulo, podrían encontrarse también en la identificación del constructo; por ejemplo, las correlaciones de una prueba de aptitud mecánica con la ejecución en cursos de oficios y en una amplia variedad de empleos contribuye a nuestra comprensión del constructo medido por la prueba. La identificación del constructo se verá además fortalecida al comparar los resultados de grupos contrastados de trabajadores que han alcanzado el éxito y con grupos que no lo logrado.

Tabla 5.2

Validación de una prueba de aritmética para diferentes propósitos

Propósito de la prueba	Pregunta ilustrativa	Evidencia de validez
Prueba de rendimiento en aritmética de la escuela elemental	¿Cuánto ha aprendido Ricardo?	Validación de contenido
Prueba de aptitud para predecir el rendimiento en matemáticas durante la secundaria	¿Qué tanto puede esperarse que aprenda Juliana?	Criterio de predicción: temporal
Técnica para diagnosticar problemas de aprendizaje	¿El rendimiento de Guillermo indica problemas específicos?	Criterio de predicción: concurrente
Medida de razonamiento cuantitativo	¿Cómo se relaciona el resultado obtenido por Elena con otros indicadores de su habilidad de razonamiento?	Identificación de constructo

Es común que en los manuales se manifieste la validez comparada con otros criterios prácticos para ayudar al usuario a comprender qué es lo que mide la prueba. Aunque tal vez no esté directamente interesado en la predicción de cualquiera de los criterios específicos empleados, al examinarlos puede hacerse una idea del área de conducta muestreada por la prueba. Si seguimos un poco más esta línea de pensamiento, vemos que todos los usos de una prueba y las interpretaciones de sus resultados implican la validez de constructo, un hecho que cada vez es más reconocido (J. P. Campbell, 1990a; Guion, 1991; Messick, 1980b, 1988, 1989; Tenopyr, 1986). Como es muy raro que las pruebas se usen en condiciones idénticas a aquellas en las que se obtuvo la validez, resulta inevitable la inclusión de cierto grado de generalización de los resultados. El significado interpretativo de los resultados de la prueba siempre se basa en constructos, que pueden variar considerablemente en su amplitud o generalización con respecto a las áreas de conducta, las poblaciones y los contextos.

Messick (1980b, 1989) argüía convincentemente que el término validez, en tanto que designa la significación interpretativa de una prueba, debería reservarse para la validez de constructo. Sostenía que otros procedimientos con los que se había asociado tradicionalmente el término "validez" tendrían que designarse con títulos más descriptivos. Así, la validez de contenido podría llamarse "relevancia del contenido" y "cobertura de contenido" podría designar a las especificaciones y representatividad de dominio, respectivamente. La validez relacionada con el criterio podría denominarse "utilidad predictiva" y "utilidad de diagnóstico" para corresponder a la validez predictiva y concurrente. Estas denominaciones más descriptivas sin duda contribuirían a una mejor comprensión de lo que realmente hacen los distintos procedimientos; de cualquier forma, en los manuales debe presentarse de manera clara la distinción entre los diferentes procedimientos de validación, ya que ello permite identificar las pruebas que se requieren para propósitos particulares.

Desde otro ángulo, incluso cuando la aplicación práctica inmediata se concentre en la descripción de contenido (como en la evaluación educativa) o la predicción del criterio (como en la selección de personal), el uso de constructos de amplitud adecuada es más efectivo que las medidas del desempeño en pruebas específicas. La investigación del criterio ha hecho cada vez más claro que tanto las medidas del criterio como los resultados de la prueba pueden expresarse de modo más efectivo como igualación de constructos. Además, la investigación de las relaciones causales entre constructos (como en el modelamiento de ecuaciones estructurales) está siendo aceptada como una contribución importante a la comprensión de cómo y por qué funcionan las pruebas.[6]

Validación en el proceso de elaboración de los tests. Cada vez se acepta más que la formulación de una prueba válida requiere de múltiples procedimientos que se emplean de manera secuenciada en diferentes etapas de la elaboración del test (Anastasi, 1986a; Guion, 1991; Jackson, 1970, 1973; N. G. Peterson *et al.*, 1990). La validez se incluye desde el inicio en la prueba en lugar de limitarse a las últimas etapas de su desarrollo, como tradicionalmente sucede en la validación relacionada con un crite-

[6] En L. A. King y D. W. King (1990) se encuentra un ejemplo de aplicación de estos procedimientos más sofisticados de validación de pruebas.

rio. El proceso de validación inicia con las definiciones detalladas del rasgo o del constructo, que se derivan de la teoría psicológica antes de emprender la investigación o la observación y el análisis sistemáticos del área de conducta relevante. Luego se preparan los reactivos que se ajusten a las definiciones del constructo, y sigue el análisis empírico de los reactivos y la selección de los más efectivos o válidos del banco inicial de reactivos. Más tarde pueden llevarse a cabo otros análisis internos apropiados, incluidos el análisis estadístico de grupos de reactivos o subtests. La etapa final comprende la validación de varias calificaciones y combinaciones interpretativas de las mismas mediante análisis estadísticos comparados con criterios externos reales.

Casi toda la información obtenida en el proceso de elaboración o uso de las pruebas es relevante para su validez. Ciertamente, los datos sobre la consistencia interna y la confiabilidad del retest ayudan a definir la homogeneidad del constructo y su estabilidad temporal. Las normas pueden proporcionar especificaciones adicionales al constructo, en especial si incluyen datos normativos separados para subgrupos clasificados por edad, sexo y otras variables demográficas que influyan en las experiencias del individuo y, por ende, en su desempeño en la prueba. Más aún, incluso después de que una prueba sale a la venta, el significado interpretativo de sus resultados puede continuar siendo aclarado y enriquecido por la acumulación gradual de observaciones clínicas y mediante proyectos especiales de investigación.[7]

Consecuencias individuales y sociales de la evaluación. Algunos psicómetras recomiendan la inclusión de un rasgo adicional en el concepto de validez de la prueba, a saber, las *consecuencias* que su aplicación tiene sobre los individuos y la sociedad. Messick (1980b, 1988, 1989, 1995) es un notable exponente de esta ampliación del concepto de validez, que pone de relieve las consecuencias no intencionales de ciertos usos de las pruebas que pueden ser nocivos para los individuos o para los miembros de ciertos grupos étnicos u otras poblaciones con diversos antecedentes. Un excelente análisis de los problemas de equilibrar diferentes metas y valores en la evaluación de solicitantes de empleo se ilustra en el informe de un comité de expertos convocado por el Consejo Nacional de Investigación (*National Research Council*), que estudió la situación con inusual minuciosidad (Hartigan y Wigdor, 1989, véanse especialmente los capítulos 13 y 14).

Las implicaciones éticas y sociales del uso de las pruebas indudablemente exigen una preocupación general. En el capítulo 18 revisaremos estos problemas, mientras que en el capítulo 6 analizaremos los aspectos más técnicos relacionados con la cuestión del "sesgo de las pruebas"; sin embargo, como han hecho notar otros psicómetras (por ejemplo, Cole y Moss, 1989), la incorporación de esas cuestiones al concepto de validez no parece ser la forma más efectiva de enfrentarlas. Son problemas a los que no se puede dar respuesta únicamente con datos empíricos y análisis estadísticos, además de que los valores deseables no deberían quedar ocultos en manipulaciones estadísticas. Necesitan ser explícitamente planteados y analizados como una meta independiente para ser considerados además de la validez estrictamente empírica y estadísticamente demostrada del uso particular de la prueba. Es posible arribar a decisio-

[7] Véase Elliott (1990b, capítulo 9), para una aplicación exitosa de este comprensivo modelo de validación de la prueba.

nes juiciosas respecto al equilibrio de metas conflictivas mediante el uso de métodos apropiados para el manejo de los sistemas de valores (Mullen y Roth, 1991; Zeichmeister y Johnson, 1992).[8] Dichos métodos requieren del juicio humano, del debate sistemático, de compromisos y de la solución de conflictos, por lo que los exponentes de los distintos sistemas deben estar adecuadamente representados. Combinar los procedimientos empíricos con apoyo estadístico para la determinación de la validez con una evaluación de las consecuencias éticas y sociales de ciertos usos particulares de las pruebas sólo confunde y oscurece la solución.

La conclusión que extraemos de la consideración de este difícil e importante problema es la reafirmación de la función crucial del usuario, que ya señalamos en el capítulo 1, y cuya responsabilidad crece cuando se requiere de juicios de valor, especialmente en los casos individuales, ya que en su elección de la prueba apropiada y en la interpretación de sus resultados puede controlar las consecuencias de la evaluación. Su orientación a los valores y su sensibilidad social contribuirán al uso adecuado de las pruebas, no sólo desde el punto de vista científico, sino también desde el ético.[9]

[8] Véase también Arkes (1993), para una revisión más amplia con referencias adicionales.

[9] Entre paréntesis, digamos que una nueva aproximación a la psicología como un todo postula "la psicología discursiva", en la que los problemas se investigan mediante el discurso de las personas en la vida diaria y los métodos experimentales tradicionales (véase, por ejemplo, Harré y Stearns, 1995; J. Smith, Harré y Van Langenhove, 1995).

Capítulo Seis

Validez: Medición e interpretación

En el capítulo 5 revisamos los conceptos de validez y las fuentes para la validación de los datos; en éste, veremos las expresiones cuantitativas de la validez y su interpretación. Los usuarios se interesan por la validez en una de sus dos etapas o en ambas. Primero, al considerar si una prueba es adecuada para sus propósitos examinan los datos que sobre su validez contiene el manual o se encuentran en otras publicaciones; con esta información se hacen una idea de las funciones psicológicas que mide y deciden si son pertinentes para el uso que pretenden dar a la prueba. En efecto, cuando los usuarios confían en los datos de validación publicados, están tratando con la validez de constructo independientemente de los procedimientos específicos que se hayan empleado para la obtención de los datos. Como vimos en el capítulo 5, no puede suponerse que los criterios empleados en los estudios publicados sean idénticos a los que el usuario desea predecir. Los puestos que en distintas compañías llevan el mismo nombre rara vez son idénticos, y dos cursos de literatura enseñados en diferentes universidades pueden ser muy distintos, de allí que al seleccionar el instrumento se requiera cierto grado de generalización de la validez.

Debido a la diversidad de las necesidades de examinación y de las inferencias que pueden derivarse de las puntuaciones de las pruebas, es posible que algunos usuarios deseen contrastar la validez del instrumento elegido con algún criterio local. Aunque los datos publicados indiquen que determinada prueba debe tener una elevada validez en una situación particular, la corroboración directa resulta deseable cuando sea técnicamente posible. La contrastación de la validez con criterios locales específicos representa la segunda etapa de la evaluación que hace el usuario. Las técnicas que estudiaremos en este capítulo son especialmente pertinentes para el análisis de los datos de validez obtenidos por los propios usuarios; sin embargo, la mayor parte también resulta útil para comprender e interpretar los datos informados en los manuales.

COEFICIENTE DE VALIDEZ Y ERROR DE ESTIMACIÓN

Medición de la relación. El coeficiente de validez es la correlación entre la puntuación de la prueba y la medida de criterio. Como proporciona un índice numérico sencillo de la validez de la prueba, suele utilizarse en los manuales para señalar la validez de una prueba comparada con cada criterio del que se tienen datos. Los datos empleados al calcular cualquier coeficiente de validez también pueden expresarse en forma de tabla o gráfica de expectancia ilustrada en el capítulo 3. De hecho, estas tablas y gráficas son un medio conveniente para mostrar lo que significa un coeficiente de validez para el evaluado. Recordará que las gráficas de expectancias muestran la probabilidad de que un individuo que obtiene cierta puntuación en la prueba obtenga un nivel especificado de desempeño en el de criterio. Veamos por ejemplo la tabla 3.6: si sabemos el resultado de un estudiante en la prueba de razonamiento numérico de los Tests de Aptitud Diferencial (*Differential Aptitude Tests*, DAT), podemos averiguar las posibilidades de que obtenga una puntuación particular en el curso de matemáticas de séptimo grado. Los mismos datos llevan a un coeficiente de validez de .60. Cuando la prueba y las variables de criterio son continuas, puede aplicarse el Coeficiente de Correlación Producto-Momento de Pearson. Cuando los datos se expresan de otra manera, como al emplear un criterio doble de aprobación-reprobación, pueden calcularse otras clases de coeficientes de correlación (véase, por ejemplo, la figura 3.7). Los procedimientos concretos para el cálculo de diferentes correlaciones pueden encontrarse en cualquier libro de estadística.

Condiciones que afectan a los coeficientes de validez. Como en el caso de la confiabilidad, resulta esencial especificar la *naturaleza del grupo* en el que se calculó el coeficiente de validez. La misma prueba puede medir diferentes funciones cuando se aplica a individuos que difieren en edad, género, nivel educativo, ocupación o en cualquier otra característica importante; por ejemplo, personas con distintos antecedentes pueden utilizar métodos de trabajo diferentes para resolver los mismos problemas de la prueba. En consecuencia, una prueba puede tener una validez elevada para predecir un criterio particular en una población y poca o nula validez en otra; o bien, puede ser una medida válida de funciones diferentes en ambas poblaciones. Las pruebas diseñadas para emplearse con diversas poblaciones deben citar en los manuales técnicos los datos apropiados sobre la posibilidad de generalizar a otras poblaciones. Más aún, en una población en la que haya grandes diferencias en las puntuaciones de la prueba, el coeficiente de validez puede diferir de manera considerable en diversas partes del rango de calificación y debe supervisarse en los subgrupos apropiados (R. Lee y Foley, 1986).

La cuestión de la *heterogeneidad de la muestra* es relevante tanto para la medición de la validez como para la medición de la confiabilidad, puesto que ambas suelen informarse en términos de coeficientes de correlación. Recordará que, en igualdad de circunstancias, entre mayor sea el rango de puntuaciones mayor será la correlación. Debe tener este hecho presente al interpretar los coeficientes de validez que aparecen en los manuales.

Una dificultad especial que se encuentra en muchas muestras de validación surge de la *preselección*; por ejemplo, una nueva prueba que se está validando para selección de personal puede aplicarse a un grupo de empleados recién contratados del que en algún momento se tendrán medidas de criterio del desempeño laboral. Sin embargo, es probable que estos empleados representen una selección superior de todos los que soli-

citaron el puesto, y, por ende, el rango del grupo en las calificaciones de la prueba y las medidas de criterio estará acortado en el extremo inferior de la distribución. En consecuencia, el efecto de la preselección consistirá en disminuir el coeficiente de validez. En el uso subsecuente de la prueba, cuando se aplica a todos los solicitantes con propósitos de selección, puede esperarse que la validez sea algo mayor.

Los coeficientes de validez también pueden cambiar con el tiempo debido a la selección de estándares cambiantes. Un ejemplo es la comparación de los coeficientes de validez calculados durante 30 años con estudiantes de Yale (Burnham, 1965). Se calcularon las correlaciones entre un indicador predictivo basado en las pruebas de la Junta Universitaria y los registros de la preparatoria, por un lado, y el promedio de calificaciones obtenido en el primer año por otro. En los 30 años la correlación cayó de .71 a .52. Un examen de las distribuciones bivariadas reveló con claridad la razón de la caída. Debido a la elevación de los estándares de admisión, las últimas generaciones fueron más homogéneas que las primeras tanto en el predictor como en la ejecución en el criterio. En consecuencia, la correlación fue menor en el último grupo, aunque la exactitud de la predicción de calificaciones individuales mostró pocos cambios. En otras palabras, la disminución observada en la correlación *no* indica que los predictores fueran menos válidos de lo que habían sido 30 años antes. Si se hubieran ignorado las diferencias en la homogeneidad del grupo, podría haberse concluido erróneamente que ése era el caso.

Para hacer una interpretación apropiada del coeficiente de validez, también debe prestarse atención a la *forma de la relación* entre la prueba y el criterio. El cálculo del coeficiente de correlación de Pearson supone que la relación es lineal y uniforme a lo largo de todo el rango. La investigación sobre la relación entre las puntuaciones de la prueba y el desempeño en el trabajo ha demostrado que generalmente se cumplen esas condiciones (Coward y Sackett, 1990; Hawk, 1970). No obstante, circunstancias especiales pueden alterar esta relación, por lo que el usuario debe estar alerta; por ejemplo, un empleo particular puede requerir un nivel mínimo de comprensión de lectura que permita a los trabajadores leer los manuales de instrucción, etiquetas y cosas similares, pero una vez que se supera este mínimo, los nuevos incrementos en la habilidad de lectura pueden no estar relacionados con el grado de éxito laboral. Éste sería un ejemplo de relación no lineal entre la prueba y el desempeño en el trabajo. Un examen de la distribución bivariada o diagrama de dispersión obtenido al graficar las puntuaciones en la lectura de comprensión contra las medidas del criterio mostrará un aumento en el desempeño laboral por encima de la mínima habilidad de lectura requerida para nivelarse después de ese punto; por consecuencia, las entradas deberían agruparse alrededor de una línea curva más que de una recta.

En otras situaciones, la línea más adecuada podría ser una recta, pero las entradas individuales alrededor de ésta pueden desviarse más en el extremo superior que en el inferior de la escala. Suponga que el desempeño en un test de aptitud académica es una condición necesaria pero no suficiente para el buen aprovechamiento en un curso. Es decir, los estudiantes con las calificaciones más bajas tendrán un mal desempeño en el curso; pero entre los estudiantes con mejores calificaciones, algunos tendrán un buen desempeño mientras que otros tendrán malas calificaciones por la poca motivación, la falta de interés u otras condiciones adversas. En esta situación, la variabilidad en el criterio de desempeño será mayor entre los estudiantes con altas puntuaciones que entre los que obtuvieron bajas calificaciones en el test. En una distribución biva-

riada se conoce a esta condición como heterosedasticidad. La correlación de Pearson supone homosedasticidad o igual variabilidad a lo largo de la distribución bivariada. En este ejemplo, la distribución bivariada tendría una forma de abanico (amplia en el extremo superior y estrecha en el inferior). Un examen de la propia distribución por lo general brinda una buena indicación de la naturaleza de la relación entre la prueba y el criterio. Las tablas y las gráficas de expectancia también revelan la eficacia relativa de la prueba en grados diferentes.

Magnitud del coeficiente de validez. ¿Qué tan elevado debe ser un coeficiente de validez? No es posible encontrar una respuesta general a esta pregunta, ya que la interpretación de un coeficiente de validez debe tener en cuenta una serie de circunstancias concomitantes. Por supuesto, la correlación obtenida debe ser lo bastante alta para ser *estadísticamente significativa* a un nivel aceptable, como los niveles de .01 y .05 que estudiamos en el capítulo 4. En otras palabras, antes de sacar conclusiones acerca de la validez de una prueba, debe tenerse una seguridad razonable de que el coeficiente de validez obtenido no se debe a fluctuaciones aleatorias del muestreo para una correlación en la población igual a cero.

Sin embargo, una vez que se ha establecido una correlación significativa entre los resultados de la prueba y el criterio, es necesario evaluar el tamaño de la correlación en función del uso que pretende hacerse de la prueba. Si se desea predecir la calificación de criterio exacta de un individuo, como el promedio de calificaciones que obtendrá en la universidad, el coeficiente de validez puede interpretarse en términos del *error estándar de estimación* (EE_{est}), que es análogo al error de medición que vimos cuando estudiamos la confiabilidad. Debe recordar que el error de medición indica el margen de error que puede esperarse en la puntuación de un individuo como resultado de la poca confiabilidad de la prueba. De modo similar, el error de estimación muestra el margen de error que puede esperarse al predecir la calificación de criterio del individuo debido a la validez imperfecta de la prueba.

El error de estimación se calcula con la siguiente fórmula:

$$EE_{est} = DE_y\sqrt{1 - r_{xy}^2}$$

en la que r_{xy}^2 es el cuadrado del coeficiente de validez y DE_y es la desviación estándar de las calificaciones del criterio. Debe notarse que si la validez fuera perfecta ($r_{xy} = 1.00$), el error de estimación sería igual a cero. Por otro lado, con una prueba que tiene una validez igual a cero, el error de estimación sería tan grande como la desviación estándar de la distribución de criterio ($EE_{est} = DE_y\sqrt{1 - 0} = DE_y$). En esas condiciones, la predicción no es mejor que una conjetura, y el margen de error de predicción es tan amplio como toda la distribución de calificaciones del criterio. Entre esos dos extremos deben encontrarse los errores de estimación correspondientes a pruebas de diversa validez.

La referencia a la fórmula para el EE_{est} mostrará que el término $\sqrt{1 - r_{xy}^2}$ sirve para indicar el tamaño del error *en relación con el error que resultaría de una mera conjetura* (es decir, con validez cero). En otras palabras, si $\sqrt{1 - r_{xy}^2}$ es igual a 1, el error de estimación es tan grande como el que encontraríamos si nos concretáramos a adivinar la calificación que obtendrá el individuo en el criterio. La mejora predictiva que puede atribuirse al uso de la prueba sería nula. Si el coeficiente de validez es .80, entonces

$\sqrt{1 - r_{xy}^2}$ es igual a .60, y el error es 60 por ciento tan grande como lo sería de manera fortuita. Dicho con otras palabras, el uso de la prueba nos permite predecir el desempeño del sujeto en el criterio con un margen de error que es 40 por ciento menor a lo que podríamos esperar si nos concretáramos a adivinar.

Parecería que incluso con una validez de .80, que es desusualmente elevada, el error de las puntuaciones predictivas es considerable. Si la función principal de las pruebas psicológicas fuera predecir la posición exacta de cada individuo en la distribución del criterio, la perspectiva sería muy desalentadora. La mayor parte de las pruebas no parecen muy eficientes cuando se examinan a la luz del error de estimación; sin embargo, en casi todas las situaciones de evaluación no es necesario predecir el desempeño de criterio específico de casos individuales, sino determinar qué individuos superarán un nivel mínimo de ejecución, o punto de corte, en el criterio. ¿Cuál es la probabilidad de que María Chaires se gradúe de la escuela de medicina, de que Tomás Silva apruebe el curso de cálculo o que Luis Domínguez tenga éxito como astronauta? ¿Qué solicitantes tienen probabilidad de convertirse en buenos oficinistas, agentes de seguros o maquinistas? La información no sólo es útil para la selección de grupos, sino también para los planes personales; por ejemplo, para una estudiante es ventajoso saber que tiene una buena oportunidad de aprobar todos sus cursos en la escuela de leyes, aunque no podamos estimar con certeza si su promedio de calificación será de 74 o de 81.

Una prueba puede mejorar considerablemente la eficacia predictiva si demuestra *cualquier* correlación significativa con el criterio, aunque sea baja. En algunas circunstancias, incluso una validez tan baja como .20 o .30 puede justificar la inclusión de la prueba en un programa de selección. Para muchos propósitos, la evaluación de las pruebas en términos del error de estimación resulta excesivamente rigurosa, por lo que resulta conveniente considerar otras formas de evaluar la contribución de una prueba que a su vez atiendan a las decisiones que deban tomarse al conocer los resultados. Veremos algunos de esos procedimientos en la siguiente sección.

VALIDEZ DE TEST Y TEORÍA DE LA DECISIÓN

Fundamentos. Supongamos que se aplica un test de aptitud a 100 solicitantes de empleo y se les sigue hasta evaluar si han tenido éxito en el trabajo asignado. La figura 6.1 muestra la distribución bivariada de los resultados de la prueba y las medidas de éxito en el trabajo de los 100 empleados. La correlación entre esas dos variables es ligeramente inferior a .70. El desempeño laboral mínimo aceptable, o punto de corte del criterio, se indica con una línea horizontal gruesa. Los 40 casos que caen por debajo de esa línea representan fracasos laborales, mientras que los 60 que están por encima obtuvieron éxito. Así, si los 100 solicitantes fueran contratados, 60 por ciento tendría éxito en el trabajo. De igual modo, si se contratara al azar a un menor número, sin contar con las puntuaciones de la prueba, la proporción de éxito probablemente sería cercana al 60 por ciento. Sin embargo, suponga que se utilizan los resultados de la prueba para seleccionar, de los 100 solicitantes, a los 45 más prometedores (razón de selección = .45). En este caso, se elegirá a los 45 individuos que caen a la derecha de la línea vertical gruesa. Puede verse que dentro de este grupo de 45 hay siete fracasos laborales, o *aceptaciones falsas*, que caen por debajo de la línea horizontal gruesa, y 38 éxitos. En

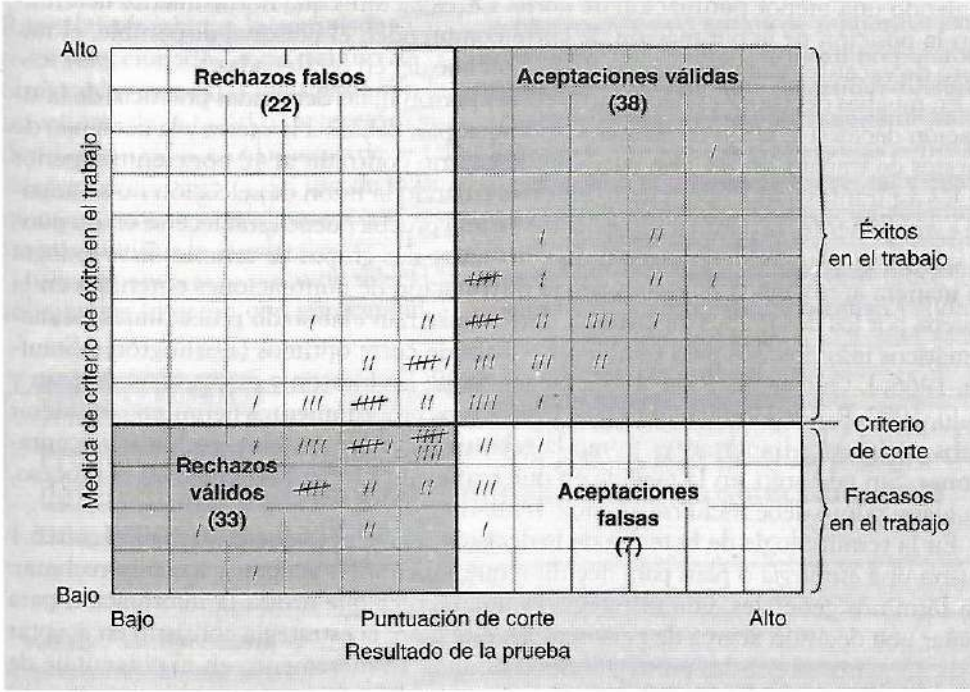


Figura 6.1. Incremento en la proporción de "éxitos" que resultan del uso de una prueba de selección.

consecuencia, el porcentaje de éxitos laborales es ahora de 84 en lugar de 60 (es decir, 38/45 = .84). Este incremento puede atribuirse al uso de la prueba como instrumento de selección. Debe hacerse notar que en la predicción de la puntuación promedio es posible ignorar los errores que no afectan la decisión. Sólo los errores de predicción que cruzan la línea de corte y por ende colocan al individuo en la categoría errónea reducen la efectividad selectiva de la prueba.

Para una completa evaluación de la efectividad de la prueba como instrumento de selección, debemos examinar en la figura 6.1 otra categoría de casos, la de *rechazos falsos*, que comprende a las 22 personas que obtuvieron una puntuación inferior a la puntuación de corte de la prueba pero superior al criterio de corte. De estos datos podemos calcular que 22 por ciento del total de la muestra de solicitantes son éxitos potenciales que se perderían de utilizar la prueba como instrumento de selección con el actual punto de corte. Al establecer la puntuación de corte debe prestarse atención al porcentaje de rechazos falsos, así como los porcentajes de éxitos y de fracasos en el grupo seleccionado. En ciertas situaciones debe establecerse un punto de corte lo bastante elevado para disminuir al mínimo la posibilidad de fracasos, lo que sería el caso si la naturaleza del trabajo es tal que un empleado mal calificado —por ejemplo, un piloto comercial— puede causar daños o pérdidas graves. En otros casos puede ser más importante admitir a tantas personas calificadas como sea posible, y correr el riesgo de incluir más fracasos. En esta situación, puede reducirse el número de falsos rechazos

eligiendo una menor puntuación de corte. Otros factores que normalmente determinan la posición de la puntuación de corte comprenden el personal disponible, el número de vacantes y la urgencia o rapidez con que deben ser llenadas.[1]

En muchas decisiones relacionadas con el personal, las demandas prácticas de la situación deciden la razón de selección; por ejemplo, debido a la oferta y la demanda de puestos de trabajo, en un caso puede ser necesario contratar al 40 por ciento superior de los aspirantes y al 75 por ciento en otro. Cuando la razón de selección no se impone externamente, la puntuación de corte de una prueba puede establecerse en un punto que permita una diferenciación máxima entre dos grupos de criterio. Esto se logra de manera aproximada si se compara la distribución de puntuaciones obtenidas en la prueba por los dos grupos de criterio. También se han elaborado procedimientos matemáticos más precisos para establecer puntos de corte óptimos (Darlington y Stauffer, 1966; I. Guttman y Raju, 1965; Jaeger, 1989; Livingston y Zieky, 1982; Martin y Raju, 1992; Rorer, Hoffman y Hsieh, 1966). Esos procedimientos permiten considerar otros parámetros importantes, como la relativa gravedad de falsos rechazos y aceptaciones. Sin embargo, en la medida en que esas evaluaciones participen en el proceso, en algún punto debe incluirse al juicio humano.

En la terminología de la teoría de la decisión, el ejemplo que presenta la figura 6.1 ilustra una *estrategia* o plan para decidir a qué solicitantes aceptar y a cuáles rechazar. En términos generales, una estrategia es una técnica que utiliza la información para tomar una decisión acerca de personas. En este caso, la estrategia consistió en aceptar a las 45 personas con las puntuaciones más altas. El incremento en el porcentaje de empleados exitosos de 60 a 84 podría usarse como base para estimar el beneficio neto del uso de la prueba.

La teoría estadística de la decisión se debe a Wald (1950), y se refería en particular a las decisiones requeridas en la inspección y control de calidad de productos industriales. Muchas de sus implicaciones para la formulación e interpretación de las pruebas psicológicas han sido investigadas sistemáticamente por Cronbach y Gleser (1965). En esencia, la teoría de la decisión es un intento por poner al proceso de decisión-elaboración en forma matemática, de modo que la información disponible pueda utilizarse para llegar a la mejor decisión en circunstancias especificadas. Algunos de los conceptos básicos de la teoría de la decisión han demostrado su utilidad en el replanteamiento y la aclaración de ciertas interrogantes sobre las pruebas. Algunas de esas ideas aparecieron en el campo de las pruebas antes que el desarrollo formal de la teoría de la decisión estadística, y más tarde se reconoció que se ajustaban a ella.

Predicción de los resultados. Las tablas de Taylor-Russell (1939), que permiten determinar la ganancia neta en la exactitud de la selección atribuible al uso de la prueba, son un antecedente de la teoría de la decisión. La información requerida incluye el coeficiente de validez de la prueba, la proporción de solicitantes que deben aceptarse (razón de selección) y la proporción de solicitantes exitosos que fueron seleccionados sin el uso de la prueba (tasa base). Un cambio en cualquiera de esas tres condiciones puede alterar la eficiencia predictiva del instrumento.

[1] En la revisión introductoria de las puntuaciones de corte en el capítulo 3 señalamos puntos similares desde un ángulo diferente.

A manera de ilustración, en la tabla 6.1 se reproduce una de las tablas de Taylor-Russell. Esta tabla es la apropiada cuando la tasa base, o porcentaje de aspirantes exitosos seleccionados antes del uso de la prueba, es de .60. Taylor y Russell prepararon otras tablas para otras tasas bases. En la parte superior de la tabla se presentan diferentes valores de la razón de selección, y en el margen se encuentran los niveles de validez. Las entradas en el cuerpo de la tabla indican la proporción de personas exitosas seleccionadas después del uso de la prueba, por lo que la diferencia entre .60 y cualquier entrada en la tabla muestra el incremento en proporción a las selecciones exitosas atribuibles a la prueba.

Obviamente, si la razón de selección fuera de 100 por ciento, es decir, si todos los solicitantes tuvieran que ser aceptados, ningún instrumento, sin importar su validez,

Tabla 6.1

Proporción de "éxitos" esperados por el uso de una prueba de determinada validez y determinada razón de selección, para una proporción base de .60

	Razón de selección										
Validez	.05	.10	.20	.30	.40	.50	.60	.70	.80	.90	.95
.00	.60	.60	.60	.60	.60	.60	.60	.60	.60	.60	.60
.05	.64	.63	.63	.62	.62	.62	.61	.61	.61	.60	.60
.10	.68	.67	.65	.64	.64	.63	.63	.62	.61	.61	.60
.15	.71	.70	.68	.67	.66	.65	.64	.63	.62	.61	.61
.20	.75	.73	.71	.69	.67	.66	.65	.64	.63	.62	.61
.25	.78	.76	.73	.71	.69	.68	.66	.65	.63	.62	.61
.30	.82	.79	.76	.73	.71	.69	.68	.66	.64	.62	.61
.35	.85	.82	.78	.75	.73	.71	.69	.67	.65	.63	.62
.40	.88	.85	.81	.78	.75	.73	.70	.68	.66	.63	.62
.45	.90	.87	.83	.80	.77	.74	.72	.69	.66	.64	.62
.50	.93	.90	.86	.82	.79	.76	.73	.70	.67	.64	.62
.55	.95	.92	.88	.84	.81	.78	.75	.71	.68	.64	.62
.60	.96	.94	.90	.87	.83	.80	.76	.73	.69	.65	.63
.65	.98	.96	.92	.89	.85	.82	.78	.74	.70	.65	.63
.70	.99	.97	.94	.91	.87	.84	.80	.75	.71	.66	.63
.75	.99	.99	.96	.93	.90	.86	.81	.77	.71	.66	.63
.80	1.00	.99	.98	.95	.92	.88	.83	.78	.72	.66	.63
.85	1.00	1.00	.99	.97	.95	.91	.86	.80	.73	.66	.63
.90	1.00	1.00	1.00	.99	.97	.94	.88	.82	.74	.67	.63
.95	1.00	1.00	1.00	1.00	.99	.97	.92	.84	.75	.67	.63
1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	.86	.75	.67	.63

Nota. En H. C. Taylor y Russell (1939) y en McCormick e Ilgen (1980, Apéndice B) puede encontrarse un conjunto completo de tablas.

(Tomado de H. C, Taylor y Russell, 1939, pág. 576)

podría mejorar el proceso de selección. La tabla 6.1 muestra que, cuando debe admitirse al 95 por ciento de los solicitantes, incluso una prueba con validez perfecta (r = 1.00) elevaría la proporción de personas exitosas apenas en tres por ciento (de .60 a .63). Por otro lado, cuando sólo es necesario elegir al cinco por ciento de los solicitantes, una prueba con un coeficiente de validez de apenas .30 puede elevar el porcentaje de solicitantes exitosos seleccionados de .60 a .82, aumento que representa el *incremento en la validez* de la prueba (Sechrest, 1963), es decir, el aumento en la validez predictiva atribuible al instrumento, que indica la contribución de la prueba a la selección de individuos que cumplan los normas mínimas en la ejecución de criterio. Al aplicar las tablas de Taylor-Russell, es obvio que la validez de la prueba debe calcularse con un grupo similar al utilizado para estimar el porcentaje de éxitos previos. En otras palabras, la contribución del instrumento no se evalúa comparándolo con el éxito aleatorio a menos que previamente se eligiera por azar a los solicitantes, circunstancia muy poco probable. Si la selección se hiciera en función de la historia laboral, cartas de recomendación y entrevistas, la contribución de la prueba tendría que evaluarse en función de lo que añade a los procedimientos anteriores.

El incremento en la validez que resulta del uso de una prueba no depende sólo de la razón de selección, sino también de la tasa base. En la situación ilustrada, la tasa base se refiere a la proporción de empleados exitosos antes de introducir la prueba con propósitos de selección. La tabla 6.1 muestra los resultados que se esperan cuando la tasa base es de .60. Para otras tasas bases necesitamos consultar la tabla apropiada en la referencia citada (H. C. Taylor y Russell, 1939). Consideremos otro ejemplo en el que la validez de la prueba es .60 y la razón de selección es de 40 por ciento. En esas condiciones, ¿cuál sería la contribución del incremento en la validez de la prueba si se empieza con una tasa base de 50 por ciento?, ¿cuál sería la contribución si empezamos con tasas bases más extremas de 10 y 90 por ciento? La referencia a la tabla Taylor-Russell apropiada para estas tasas muestra que el porcentaje de empleados exitosos aumenta de 50 a 75 por ciento en el primer caso; de 10 a 21 por ciento en el segundo, y de 90 a 99 por ciento en el tercero. Así, la mejora en el porcentaje de empleados exitosos atribuible al uso de la prueba es de 25 cuando la tasa base era de 50, pero de sólo 11 y nueve cuando las tasas bases eran más extremas.

Las implicaciones de las tasas bases extremas son de especial interés en la psicología clínica, en la que la tasa base se refiere a la frecuencia de la condición patológica que debe diagnosticarse en la población examinada (Buchwald, 1965; Cureton, 1957a; Meehl y Rosen, 1955; J. S. Wiggins, 1973/1988). Por ejemplo, si cinco por ciento de la población admitida de una clínica padece daño orgánico cerebral, entonces en esa población la tasa base del daño cerebral es de cinco por ciento. Aunque la introducción de cualquier prueba válida mejore la precisión predictiva o de diagnóstico, la mejora es mayor cuando las tasas bases se aproximan a 50 por ciento; sin embargo, con las tasas bases extremas que se encuentran en las condiciones patológicas raras, la mejora puede ser insignificante. En estas condiciones, el uso de una prueba puede resultar inadecuado cuando se considera el costo de su aplicación y calificación. En una situación clínica, el costo puede incluir el tiempo del personal que podría dedicarse a la atención de otros casos (Buchwald, 1965). El número de falsos positivos, o individuos normales clasificados incorrectamente como patológicos, aumentaría desde luego el costo global en una situación clínica.

Cuando la gravedad de una condición rara hace que su diagnóstico sea urgente, pueden emplearse pruebas de validez moderada en la primera etapa de una secuencia de decisiones; por ejemplo, se puede empezar por elegir todos los casos con una prueba de fácil aplicación y validez moderada. Si la puntuación de corte se hace lo suficientemente alta (cuando son favorables los resultados elevados), habrá unos cuantos falsos negativos pero muchos falsos positivos o sujetos normales diagnosticados como patológicos. Los últimos pueden detectarse luego con exámenes individuales más intensivos aplicados a todos los casos diagnosticados como positivos por la prueba. Esta solución podría ser apropiada, por ejemplo, cuando las condiciones hacen que la evaluación individual intensiva de todos los casos resulte imposible.

Relación entre validez y productividad. En muchas situaciones prácticas se desea estimar el efecto de la prueba de selección, no el porcentaje de quienes superan el desempeño mínimo, sino la productividad global de los seleccionados. ¿Cómo es el nivel actual de eficiencia o desempeño en el criterio de los trabajadores contratados sobre la base de la prueba en comparación con la de la muestra total de solicitantes que habrían sido contratados sin la aplicación del instrumento? Varios investigadores abordaron la cuestión siguiendo el trabajo de Taylor y Russell. Brogden (1946b) demostró primero que el incremento esperado en el resultado es directamente proporcional a la validez de la prueba. Así, la mejora que resulta de usar una prueba con validez de .50 es 50 por ciento tan grande como la que se espera de una prueba de validez perfecta.

La relación entre la validez de la prueba y el aumento esperado en el desempeño en el criterio puede verse en la tabla 6.2.[2] Al expresar las puntuaciones del criterio como puntuaciones estándares con una media de cero y una *DE* igual a 1.00, la tabla proporciona la puntuación media del criterio que se espera de trabajadores seleccionados con una prueba de determinada validez y con cierta razón de selección. En este contexto, el resultado de base promedio, que incumbe al desempeño de los solicitantes seleccionados sin usar la prueba, se presenta en la columna correspondiente a la validez cero (usar una prueba con validez cero equivale a no emplear prueba alguna). Para ilustrar el uso de la tabla, supongamos que se contrata al 20 por ciento de los solicitantes con mayor puntuación (razón de selección = .20) mediante una prueba cuyo coeficiente de validez es .50. La referencia a la tabla 6.2 muestra que la ejecución media del criterio de este grupo es .70 *DE* encima de la base media esperada de una muestra no examinada. Con la misma razón de selección de 20 por ciento y una prueba perfecta (coeficiente de validez = 1.00), la calificación del criterio promedio de los solicitantes aceptados sería de 1.40, justo el doble de lo que sería con la prueba de validez de .50. Si se comparan otros desempeños del criterio promedio en cualquier hilera de la tabla 6.2, se encontrarán relaciones lineales similares; por ejemplo, con una razón de selección de 60 por ciento, una validez de .25 arroja un resultado promedio de .16, mientras que una validez de .50 un resultado medio de .32. Una vez más, duplicar la validez implica hacer lo mismo con el incremento en el resultado.

Schmidt y sus colaboradores (Schmidt, Hunter, McKenzie y Muldrow, 1979) realizaron un análisis de la productividad en relación con la validez de la prueba. Usando

[2] Naylor y Shine (1965) prepararon una tabla que incluye más valores para las razones de selección y los coeficientes de validez.

Tabla 6.2

Calificación media estándar de la puntuación de criterio de los casos aceptados en relación con la validez de la prueba y la razón de selección.

Razón de selección	Coeficiente de validez																			
	.00	.05	.10	.15	.20	.25	.30	.35	.40	.45	.50	.55	.60	.65	.70	.75	.80	.85	.90	.95	1.00
.05	.00	.10	.21	.31	.42	.52	.62	.73	.83	.94	1.04	1.14	1.25	1.35	1.46	1.56	1.66	1.77	1.87	1.98	2.08
.10	.00	.09	.18	.26	.35	.44	.53	.62	.70	.79	.88	.97	1.05	1.14	1.23	1.32	1.41	1.49	1.58	1.67	1.76
.15	.00	.08	.15	.23	.31	.39	.46	.54	.62	.70	.77	.85	.93	1.01	1.08	1.16	1.24	1.32	1.39	1.47	1.55
.20	.00	.07	.14	.21	.28	.35	.42	.49	.56	.63	.70	.77	.84	.91	.98	1.05	1.12	1.19	1.26	1.33	1.40
.25	.00	.06	.13	.19	.25	.32	.38	.44	.51	.57	.63	.70	.76	.82	.89	.95	1.01	1.08	1.14	1.20	1.27
.30	.00	.06	.12	.17	.23	.29	.35	.40	.46	.52	.58	.64	.69	.75	.81	.87	.92	.98	1.04	1.10	1.16
.35	.00	.05	.11	.16	.21	.26	.32	.37	.42	.48	.53	.58	.63	.69	.74	.79	.84	.90	.95	1.00	1.06
.40	.00	.05	.10	.15	.19	.24	.29	.34	.39	.44	.48	.53	.58	.63	.68	.73	.77	.82	.87	.92	.97
.45	.00	.04	.09	.13	.18	.22	.26	.31	.35	.40	.44	.48	.53	.57	.62	.66	.70	.75	.79	.84	.88
.50	.00	.04	.08	.12	.16	.20	.24	.28	.32	.36	.40	.44	.48	.52	.56	.60	.64	.68	.72	.76	.80
.55	.00	.04	.07	.11	.14	.18	.22	.25	.29	.32	.36	.40	.43	.47	.50	.54	.58	.61	.65	.68	.72
.60	.00	.03	.06	.10	.13	.16	.19	.23	.26	.29	.32	.35	.39	.42	.45	.48	.52	.55	.58	.61	.64
.65	.00	.03	.06	.09	.11	.14	.17	.20	.23	.26	.28	.31	.34	.37	.40	.43	.46	.48	.51	.54	.57
.70	.00	.02	.05	.07	.10	.12	.15	.17	.20	.22	.25	.27	.30	.32	.35	.37	.40	.42	.45	.47	.50
.75	.00	.02	.04	.06	.08	.11	.13	.15	.17	.19	.21	.23	.25	.27	.30	.32	.33	.36	.38	.40	.42
.80	.00	.02	.04	.05	.07	.09	.11	.12	.14	.16	.18	.19	.21	.22	.25	.26	.28	.30	.32	.33	.35
.85	.00	.01	.03	.04	.05	.07	.08	.10	.11	.12	.14	.15	.16	.18	.19	.20	.22	.23	.25	.26	.27
.90	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09	.10	.11	.12	.13	.14	.15	.16	.17	.18	.19	.20
.95	.00	.01	.01	.02	.02	.03	.03	.04	.04	.05	.05	.06	.07	.07	.08	.08	.09	.09	.10	.10	.11

(Tomado de Brown y Ghiselli, 1953, pág. 342).

como ejemplo el puesto de programador de computadora en el gobierno federal estadounidense, investigaron el valor económico del incremento de la productividad observado después de usar durante un año una prueba de aptitud para el uso de las computadoras (validez = .76) en las nuevas contrataciones. Llegaron a sus estimados por la aplicación de las técnicas de la teoría de la decisión a los datos disponibles en la Oficina de Manejo de Personal de los Estados Unidos. Se calcularon las ganancias esperadas para nueve razones de selección que iban de .50 a .80 y para cinco coeficientes de validez de procedimientos previos de selección que iban de cero (selección al azar) a .50.

Los resultados indicaron ganancias impresionantes en la productividad desde que empezó a utilizarse la prueba en todas estas condiciones. Cuando se comparó el uso de la prueba con una selección al azar, la ganancia en dólares fluctuó entre 97.2 millones para una razón de selección de .05 a 16.5 millones para una razón de selección de .80. Con la validez de selección previa de .50, las ganancias correspondientes fluctuaron entre 33.3 y 5.6 millones de dólares. Estas ganancias se extendieron durante la permanencia esperada de los empleados recién contratados, que para los programadores de computadora del gobierno federal estadounidense es en promedio ligeramente superior a 10 años. Debe mencionarse que los estimados se basan en la suposición de que los procedimientos de selección inician con los solicitantes de mayor calificación y avanzan en orden descendente hasta que se alcanza la razón de selección especificada. En otras palabras, el procedimiento supone un uso óptimo del proceso de selección.

Usando los datos del censo para evaluar el número de programadores de cómputo en la población total de los Estados Unidos, los investigadores también obtuvieron los estimados correspondientes del efecto de emplear la prueba a nivel nacional. En un estudio subsecuente, Hunter y Schmidt (1981) exploraron la posible aplicación de las mismas técnicas estadísticas a la fuerza laboral nacional en todas las ocupaciones. Las estimaciones preliminares son declaradamente crudas y tentativas, y los procedimientos alternativos tienden a arrojar estimaciones bajas (Burke y Frederick, 1984; U.S. Department of Labor, 1983b; Weekley, Frank, O'Connor y Peters, 1985), pero los datos disponibles indican que el uso de métodos efectivos para asignar los puestos puede contribuir sustancialmente a incrementar la productividad. Cada vez es mayor la atención que se concede a la naturaleza de la productividad en el trabajo, así como a las condiciones individuales y organizacionales que la afectan. Se trata de un área de investigación de criterios en la validación de las pruebas que muestra grandes avances teóricos y metodológicos (J. P. Campbell, Campbell & Associates, 1988; Hunter, Schmidt y Judiesch, 1990; Raju, Burke y Normand, 1990).

El concepto de utilidad en la teoría de la decisión. Una característica de la teoría de la decisión es que las pruebas se evalúan en términos de su eficacia en una situación particular. La evaluación no sólo toma en consideración la validez de la prueba para predecir un criterio particular, sino también otros parámetros, como la tasa base y la razón de selección. Otro parámetro importante es la relativa *utilidad* de los resultados esperados, lo favorable o desfavorable que se considere cada resultado. La falta de sistemas adecuados para asignar valores a los resultados en términos de una escala uniforme de utilidad ha sido uno de los principales obstáculos a la aplicación de la teoría de la decisión. En las decisiones industriales, es frecuente asignar un valor económico a diferentes resultados, pero incluso en esos casos es difícil evaluar en términos monetarios ciertos resultados relacionados con el bienestar público, las relaciones públicas y la mo-

ral de los empleados. Las decisiones educativas deben tener en cuenta las metas institucionales, los valores sociales y otros factores relativamente intangibles. Las decisiones individuales, como en la consejería, deben considerar las preferencias y el sistema de valores del individuo. Sin embargo, se ha señalado repetidamente que la teoría de la decisión no introduce al proceso de decisión el problema de los valores, sino que se concreta a hacerlo explícito. Los sistemas de valores siempre han estado presentes en las decisiones, pero hasta ahora no eran reconocidos o manejados en forma sistemática.

La investigación realizada por Schmidt, Hunter y colaboradores, citada en la sección precedente, ilustra los avances en los procedimientos para la asignación de valores en los modelos de toma de decisiones. Aunque se interesan en el valor económico de los bienes y servicios proporcionados por los trabajadores, las técnicas desarrolladas en dicha investigación también pueden aplicarse a la medición de otros valores. Al basarse en la cuantificación de los juicios humanos, los mismos procedimientos pueden usarse con cualquier escala numérica arbitraria, siempre que se defina con claridad y se aplique de manera consistente. Debe hacerse notar que las estimaciones requeridas por los modelos de toma de decisiones sólo tienen que ver con los valores relativos, no los absolutos, de diferentes resultados. En Boudreau (1991) puede encontrarse un tratamiento técnico de la utilidad en la toma de decisiones sobre el personal.[3]

Al elegir una estrategia de decisión, la meta es aumentar al máximo las ganancias esperadas de todos los resultados. La representación esquemática de una estrategia simple de decisión que se observa en la figura 6.2 ayudará a hacer más claro el procedimiento. El diagrama muestra la estrategia de decisión ilustrada en la figura 6.1, en la que se aplica una sola prueba a un grupo de solicitantes y la decisión de aceptar o rechazar a cada uno se toma sobre la base de la puntuación de corte de la prueba. Hay cuatro resultados posibles, incluyendo las aceptaciones válidas y falsas y los rechazos válidos y falsos. La probabilidad de cada resultado puede encontrarse del número de personas en cada una de las cuatro secciones de la figura 6.1. Como en el ejemplo había 100 solicitantes, esos números divididos entre 100 proporcionan las probabilidades de los cuatro resultados presentados en la figura 6.2.

Los otros datos requeridos son las utilidades de los distintos resultados, expresados en una escala común. Estos valores hipotéticos, cualquiera que sea el procedimiento de juicio por el que se obtuvieron, se presentan en la última columna de la figura 6.2. Para encontrar la utilidad global esperada de la estrategia se multiplica la probabilidad de cada resultado por su utilidad, se suman estos productos para las cuatro calificaciones y se resta una cifra correspondiente al costo de la examinación. Este último término pone de relieve el hecho de que es más probable que se conserve una prueba de poca validez si es breve, barata, de fácil aplicación por personal con poco entrenamiento y adecuada para su aplicación colectiva. Una prueba individual que requiere de un examinador capacitado o de equipo costoso necesita una mayor validez para justificar su uso. En el ejemplo hipotético de la figura 6.2, el costo de la examinación se estima como .10 en la escala de utilidad. La utilidad total esperada (*UE*) para esta estrategia de decisión es:

$$UE = (.38)(1.00) + (.07)(-1.00) + (.33)(0) + (.22)(-.50) - .10 = +.10$$

[3] Para otro punto de vista de la evaluación de la utilidad, véase Cascio y Morris (1990), Messick (1989, pp. 78–81), y Sadacca, Campbell, Difazio, Schultz y White (1990).

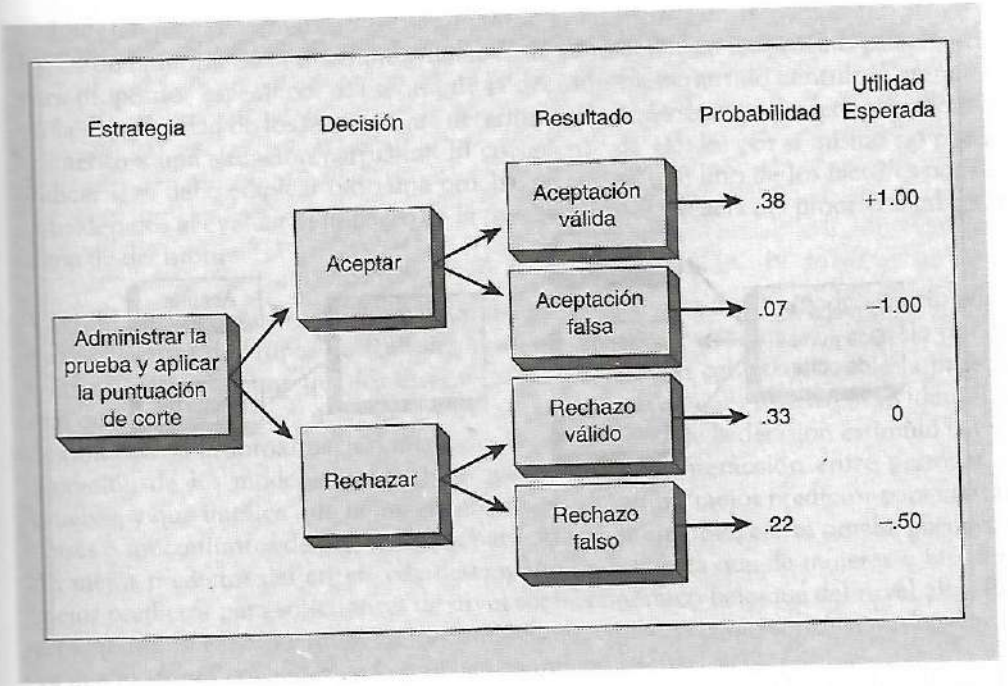


Figura 6.2. Una estrategia simple de decisión.

Esta *UE* puede compararse con otras *UE* encontradas con diferentes puntos de corte, con distintas pruebas (que difieren en validez y costo) o con una batería de pruebas, así como con diferentes estrategias de decisiones.[4]

Estrategias secuenciales y tratamientos adaptativos. En algunas situaciones, la eficacia de la prueba puede aumentar con el uso de estrategias más complejas de decisión que tienen en consideración otros parámetros. Dos ejemplos servirán para ilustrar esas posibilidades. Primero, las pruebas pueden servir para tomar decisiones *secuenciales* más que terminales. Con la simple estrategia de decisión ilustrada en las figuras 6.1 y 6.2, todas las decisiones de aceptación o rechazo se consideran terminales. La figura 6.3, por otro lado, muestra una decisión en una secuencia de dos etapas. La prueba A puede ser un instrumento de selección breve y de fácil aplicación. Tomando como base el desempeño en esta prueba, los individuos se distribuyen en tres categorías, las de los claramente aceptados o rechazados así como el grupo intermedio de casos "inciertos" que debe ser examinado con técnicas más intensivas, representadas por la prueba B. A partir del examen aplicado en la segunda etapa, los individuos se asignarán a la categoría de aceptados o de rechazados.

Otra estrategia, adecuada para el diagnóstico de los trastornos psicológicos, consiste en usar sólo dos categorías y probar luego a *todos* los casos clasificados como positi-

[4] Véase J. S. Wiggins (1973/1988, cap. 6) para encontrar ejemplos de varias estrategias de decisión que muestran paso a paso todos los cálculos.

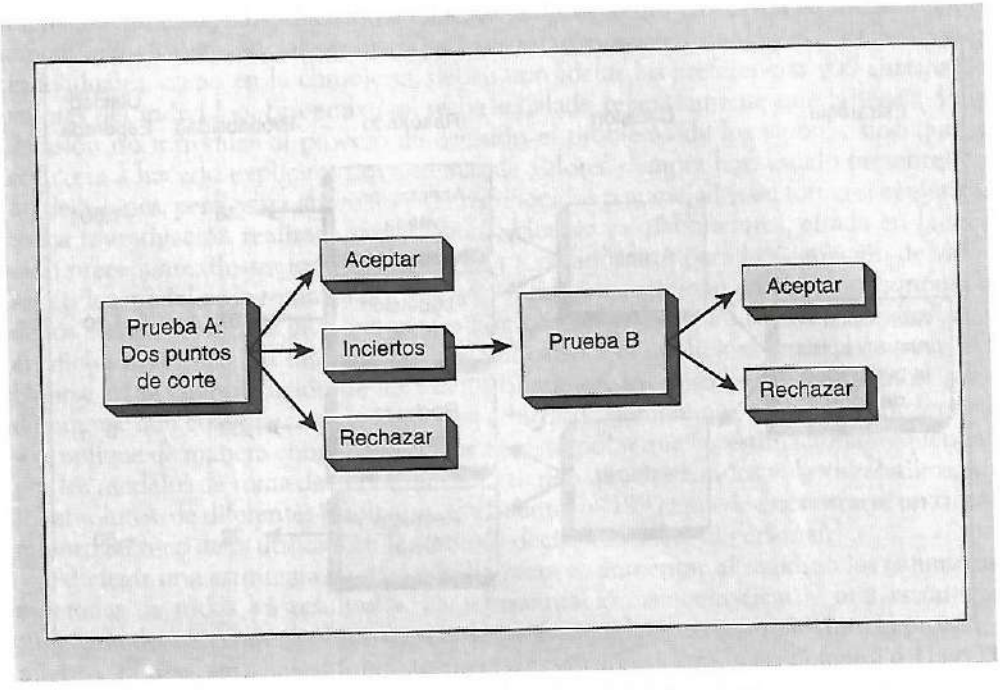


Figura 6.3. Una estrategia secuencial de decisión.

vos (es decir, posiblemente patológicos) en la prueba preliminar de selección. Ésta es la estrategia que citamos en relación con el uso de las pruebas para diagnosticar condiciones patológicas con tasas bases muy bajas.

Debe hacerse notar que de hecho muchas de las decisiones relacionadas con el personal son secuenciales, aunque tal vez no se perciban así. Los empleados incompetentes contratados por errores de predicción por lo general pueden ser despedidos luego de un periodo de prueba; los estudiantes fracasados pueden ser expulsados de la universidad en diversas etapas. En tales situaciones, son sólo las decisiones adversas las terminales. Es indudable que las decisiones de selección incorrectas que se rectifican más tarde pueden ser costosas en términos de varios sistemas de valores, pero suelen serlo menos que las decisiones terminales erróneas.

Otra condición que puede alterar la eficacia de una prueba psicológica es la disponibilidad de tratamientos alternativos y la posibilidad de *tratamientos adaptados* a características individuales. Un ejemplo sería la utilización de distintos procedimientos de entrenamiento para trabajadores con diferentes niveles de aptitud o la introducción de programas de remedio para estudiantes con ciertos problemas educativos. En estas condiciones, la estrategia de decisión seguida en cada caso debe considerar los datos disponibles sobre la interacción de la puntuación inicial de la prueba y el tratamiento diferencial. Cuando se emplean tratamientos adaptados a las necesidades individuales, es probable que la tasa de éxito aumente considerablemente. Como la asignación de individuos a tratamientos alternativos es en esencia un problema de clasificación más que de selección, en la sección correspondiente diremos más acerca de la metodología requerida.

Los ejemplos presentados ilustran algunas de las formas en que los conceptos y la lógica de la teoría de la decisión ayudan en la evaluación de las pruebas psicológicas para propósitos específicos. La teoría de la decisión ha permitido centrar la atención en la complejidad de los factores que determinan la contribución de determinado instrumento a una situación particular. El coeficiente de validez por sí mismo no puede indicar si se debe emplear o no una prueba, ya que es sólo uno de los factores por ser considerados al evaluar el impacto de la prueba sobre la eficacia del proceso total de la toma de decisiones.[5]

Variables moderadoras. La validez de una prueba para determinado criterio puede variar entre subgrupos que difieren en características personales. El modelo psicométrico clásico supone que los errores de predicción son característica de la prueba más que de la persona, y que esos errores se distribuyen al azar entre los individuos. La flexibilidad de la aproximación introducida por la teoría de la decisión estimuló la exploración de los modelos predictivos que incluían la interacción entre personas y pruebas, y que implica que la misma prueba puede ser un mejor predictor para ciertas clases o subconjuntos de personas que para otras; por ejemplo, cierta prueba puede ser un mejor predictor del criterio de desempeño de hombres que de mujeres o bien un mejor predictor para solicitantes de nivel socioeconómico bajo que del nivel alto. En esos ejemplos, género y nivel socioeconómico se conocen como *variables moderadoras*, ya que moderan la validez de la prueba (Saunders, 1956).

Los intereses y la motivación pueden funcionar como variables moderadoras, de modo que si un solicitante tiene poco interés en su trabajo, su desempeño será probablemente malo cualquiera que sea la puntuación que haya obtenido en los tests de aptitud. Entre estas personas, la correlación entre las puntuaciones de la prueba de aptitud y el desempeño laboral será baja. Por otro lado, para los individuos que estén interesados y motivados, la correlación entre la puntuación de la prueba de aptitud y el éxito laboral será muy alta. Las décadas de los cincuenta y sesenta presenciaron una racha de investigaciones sobre una amplia variedad de posibles variables moderadoras. Una serie de estudios de Ghiselli (1956, 1960, 1963, 1968) abordó la predicción del desempeño laboral. Otros investigadores sometieron a prueba hipótesis sobre el papel de las variables de personalidad, particularmente en la predicción de aprovechamiento académico (N. Frederiksen y Gilbert, 1960; N. Frederiksen y Melville, 1954; Grooms y Endler, 1960; L. J. Stricker, 1966).

Un descubrimiento relativamente constante fue una diferencia de género en la predicción de grados académicos. Estudios de muchas fuentes que cubrían varios cientos de coeficientes de correlación informaron correlaciones mayores para las mujeres que para los hombres entre las puntuaciones de los tests de aptitud y los grados académicos (Gross, Faggen y McCarthy, 1974; Schmitt, Mellon y Bylenga, 1978; Seashore, 1962). Se descubrió la misma tendencia en la educación media y la superior, aunque fue más pronunciada a nivel universitario. Los datos no indican la razón de esta diferencia de género en lo predecible del rendimiento académico, pero es interesante es-

[5] Véase J. S. Wiggins (1973/1988, cap. 6) para una revisión completa de las implicaciones de la teoría de la decisión para el uso de las pruebas, Cronbach y Gleser (1965) para un nivel más técnico.

pecular al respecto considerando otras diferencias de género conocidas. Si las estudiantes en general tienden a ser más conformistas y más inclinadas a aceptar los valores y las normas de la situación escolar, su rendimiento probablemente dependerá en gran medida de sus habilidades. Por otro lado, si los estudiantes tienden a concentrar sus esfuerzos en las actividades (extraescolares) que despiertan su curiosidad, estas diferencias de interés pueden introducir una varianza adicional en su rendimiento académico y hacer más difícil la predicción de su aprovechamiento a partir de las puntuaciones obtenidas en las pruebas de aptitud. Sin embargo, observe que las diferencias de género en estos coeficientes de validez, aunque constantes, generalmente son pequeñas. Más aún, existen indicios de que las diferencias tendían a ser más pequeñas en los estudios más recientes, un descubrimiento que podría reflejar un cambio en las actitudes de las mujeres de finales de los sesenta y los setenta.

En general, las expectativas sobre la contribución de las variables moderadoras no se cumplieron (Abrahams y Alf, 1972; Pinder, 1973; Zedeck, 1971). Los análisis metodológicos revelaron muchas fallas; muchas veces la validación cruzada en nuevas muestras no corroboró los descubrimientos iniciales y no parecía probable que el uso de moderadores mejorara sustancialmente la predicción que podría lograrse por otros medios. A la luz de los conocimientos actuales, si no se cuenta con evidencias explícitas, no puede *suponerse* que variable alguna modere la validez. Con todo, el concepto de variables moderadoras puede tener valor heurístico para aumentar la comprensión de la conducta individual (como en los estudios de caso clínicos) y para proponer nuevas hipótesis que deben investigarse con los controles metodológicos adecuados. De hecho, las décadas de los ochenta y los noventa atestiguaron el resurgimiento del interés por las variables moderadoras. Aunque no fueron detectadas por los primeros procedimientos, algunas variables moderadoras pueden ser ahora identificadas mediante análisis estadísticos más elaborados (Morris, Sherman y Mansfield, 1986; E. F. Stone y Hollenbeck, 1989).

COMBINACIÓN DE INFORMACIÓN A PARTIR DE DIFERENTES TESTS

Para la predicción de criterios prácticos, a menudo no se requiere una sino varias pruebas. Los criterios son, en su mayor parte, complejos, y la medida de criterio depende de varios rasgos diferentes. Si se diseñara una prueba para medir este criterio tendría que ser muy heterogénea, pero ya hemos dicho que es más conveniente una prueba relativamente homogénea (que mida más bien un solo rasgo) porque produce puntuaciones menos ambiguas (capítulo 5). Por ende, a menudo es preferible usar una combinación de pruebas relativamente homogéneas, cada una de las cuales cubra un aspecto diferente del criterio, en lugar de aplicar una sola consistente en una mescolanza de muchas tipos distintos de reactivos.

A las pruebas que se emplean en una serie especialmente seleccionadas para predecir un solo criterio se las conoce como *batería de pruebas*. El problema principal del uso de tales baterías tiene que ver con la forma en que se combinan las puntuaciones de estos distintos instrumentos para llegar a una decisión con respecto a cada individuo. Con este propósito se siguen procedimientos que caen en dos categorías principales: la ecuación de regresión múltiple y el análisis de perfiles. Cuando las pruebas se aplican en el estudio intensivo de casos individuales, como en el diagnóstico clínico, la consejería o la evalua-

ción de ejecutivos de alto nivel, es común que el examinador utilice los resultados de las pruebas sin mayor análisis estadístico. Al preparar el informe del caso y hacer las recomendaciones, el examinador confía en su juicio, experiencia y bases teóricas para interpretar los patrones de puntuaciones e integrar los descubrimientos de diferentes pruebas.

Ecuación de regresión múltiple. La ecuación de regresión múltiple predice una puntuación de criterio para cada individuo sobre la base de sus puntuaciones en todas las pruebas de la batería. La siguiente ecuación de regresión múltiple ilustra la aplicación de esta técnica para predecir el aprovechamiento de un estudiante en los cursos de matemáticas en la secundaria a partir de las puntuaciones que haya obtenido en las pruebas verbal (V), numérica (N) y de razonamiento (R):

$$\text{Aprovechamiento en matemáticas} = .21V + .21N + .32R + 1.35$$

En este ejemplo, los resultados de las pruebas y la puntuación del criterio se expresan como estaninas, aunque puede emplearse cualquier otra escala de calificación para este propósito. En la ecuación anterior, la calificación estanina del estudiante en cada una de las tres pruebas se multiplica por el correspondiente peso dado en la ecuación. La suma de esos productos, más una constante (1.35), proporciona la posición estanina del estudiante en el curso de matemáticas.

Suponga que Beatriz Samaniego obtuvo las siguientes calificaciones estanina:

Verbal	6
Numérica	4
Razonamiento	8

El aprovechamiento estimado en matemáticas para esta estudiante se encuentra de la siguiente manera:

$$\text{Aprovechamiento en matemáticas} = (.21)(6) + (.21)(4) + (.32)(8) + 1.35 = 6.01$$

La estanina que se predice para Beatriz es de aproximadamente 6. Se recordará (capítulo 3) que una estanina de 5 representa un desempeño promedio, por lo que puede esperarse que Beatriz tenga un aprovechamiento mejor que el promedio en sus cursos de matemáticas. Su desempeño superior en la prueba de razonamiento ($R = 8$), y su calificación por encima del promedio en la prueba verbal ($V = 6$) compensan su mediocre calificación en la velocidad y precisión para hacer cálculos ($N = 4$).

En los textos de estadísticas aplicadas a la psicología pueden encontrarse técnicas específicas para el cálculo de las ecuaciones de regresión (por ejemplo, D. C. Howell, 1997; Runyon y Haber, 1991). En esencia, la ecuación se basa en la correlación de cada prueba con el criterio, así como en las intercorrelaciones de las pruebas. Desde luego, las pruebas que correlacionan más alto con el criterio deben recibir más peso, pero es igualmente importante tomar en consideración la correlación de cada prueba con las

otras que componen la batería. Las pruebas que correlacionan alto entre sí representan una duplicación innecesaria, ya que cubren en gran medida los mismos aspectos del criterio. La inclusión de dos de esas pruebas no aumentará en forma apreciable la validez de la batería, aunque ambas puedan correlacionar bien con el criterio. En dicho caso, una de ellas puede ser tan eficaz como el par, por lo que sólo una debería mantenerse en la batería.

Pero aun después de que se han eliminado los casos más serios de duplicación, las pruebas restantes en la batería correlacionarán entre sí en diversos grados. Para obtener un valor predictivo máximo, las pruebas que hacen una contribución casi única a la batería deberían recibir mayor peso que las que duplican en parte las funciones de otras pruebas. En el cálculo de una ecuación de regresión múltiple, cada una recibe un peso en proporción directa a su correlación con el criterio y en proporción inversa a su correlación con las otras pruebas. Así, el mayor peso se asignará a la prueba con la mayor validez y el menor traslape con el resto de la batería.

La validez de la batería puede encontrarse calculando la correlación múltiple (R) entre el criterio y la batería. Esta correlación indica el mayor valor predictivo que puede obtenerse de una batería dada cuando cada prueba recibe un peso óptimo por predecir el criterio en cuestión. Los pesos óptimos son los que determina la ecuación de regresión.

Observe que esos pesos son óptimos sólo para la muestra particular en que se encontraron. Debido a los errores aleatorios en los coeficientes de correlación usados al derivarlos, los pesos de regresión pueden variar de una muestra a otra, por lo que la batería debe validarse de manera cruzada correlacionando las puntuaciones del criterio que predice con los resultados del criterio reales en una nueva muestra. Aunque se dispone de fórmulas que permiten estimar la cantidad de la *disminución* esperada en una correlación múltiple cuando se aplica la ecuación de regresión a una segunda muestra, siempre que sea posible es preferible realizar la verificación empírica. Entre mayor sea la muestra de la que se derivaron los pesos de regresión, menor será la disminución o pérdida.[6]

En ciertas situaciones, la validez predictiva de la batería aumenta si se incluye en la ecuación de regresión una prueba que tenga una correlación cero con el criterio pero una correlación elevada con otra prueba de la batería. Esta curiosa situación surge cuando la prueba no correlacionada con el criterio actúa como *variable de supresión* para eliminar o suprimir la varianza irrelevante en la otra prueba (Conger y Jackson, 1972). Por ejemplo, la comprensión de lectura puede correlacionar alto con las puntuaciones de una prueba de aptitud matemática o mecánica, porque los problemas de la prueba requieren la comprensión de complicadas instrucciones escritas. Si la comprensión de lectura es irrelevante para la conducta laboral que pretende predecirse, la que requiere la prueba introduce varianza de error y disminuye la validez predictiva de la prueba. Aplicar una prueba de comprensión de lectura e incluir sus puntuaciones en la ecuación de regresión eliminará esta varianza de error al aumentar la validez de la batería. La variable supresora aparece en la ecuación de regresión con un peso negativo, de modo que entre mayor sea la puntuación que un individuo obtenga en lectura de comprensión, más es lo que se descuenta de su resultado en la prueba de matemáticas o de mecánica. Con todo, en cualquier situación es preferible el procedi-

[6] En algunas circunstancias, los pesos de unidad u otras alternativas pueden ser preferibles a los pesos de regresión. Para una breve revisión de la investigación sobre diversos esquemas de pesos, véase Dunnette y Borman (1979).

miento más directo de revisar una prueba para eliminar la varianza irrelevante a la eliminación estadística indirecta mediante una variable supresora. Cuando no son factibles los cambios en la prueba, debe considerarse la investigación de las variables de supresión, pero su efecto siempre debe verificarse en una nueva muestra.

Análisis de perfiles y puntuaciones de corte. Además del análisis individualizado de perfiles empleado en la evaluación clínica, el patrón de puntuaciones obtenidas con una batería de selección de personal puede evaluarse en términos de puntuaciones múltiples de corte. En pocas palabras, el procedimiento consiste en establecer para cada prueba una puntuación mínima de corte. Cuando se aplica estrictamente, se rechaza a todo el que cae por debajo de la puntuación mínima de *cualquiera* de las pruebas apropiadas. Al elegir las pruebas apropiadas y establecer las puntuaciones de corte para una determinada ocupación, se acostumbra considerar algo más que la validez de las pruebas. Si sólo tomaran en consideración las pruebas que arrojaran coeficientes de validez significativos, podrían pasarse por alto una o más habilidades esenciales en que todos los trabajadores sobresalen, de ahí la necesidad de considerar también las aptitudes en las que los trabajadores destacan como grupo, incluso cuando las diferencias individuales luego de cierto mínimo no estén relacionadas con el grado de éxito laboral. Más aún, en algunos empleos los trabajadores pueden ser tan homogéneos en un rasgo clave que el rango de las diferencias individuales es demasiado estrecho para arrojar una correlación significativa entre la puntuación de la prueba y el criterio.

El uso del método múltiple de corte lo ilustra la Batería de Pruebas de Aptitudes Generales, BTAG (*General Aptitude Test Battery, GATB*)[7] desarrollada por el Servicio de Empleo de los Estados Unidos para usarse en la consejería ocupacional y programa de referencia de sus oficinas estatales de servicio para el empleo (U.S. Department of Labor, 1970). De las nueve puntuaciones de aptitud obtenidas con esta batería, se eligió a las que serían consideradas para cada ocupación sobre la base de las correlaciones del criterio, así como las medias y desviaciones estándares de los trabajadores en esa ocupación y las observaciones cualitativas de los analistas de puestos.

El argumento más sólido a favor del uso de puntuaciones múltiples de corte en lugar de una ecuación de regresión se basa en la posibilidad de la intervención de puntuaciones compensatorias. En otras palabras, en la puntuación total de la batería puede pasar inadvertida una deficiencia grave si el sujeto obtiene una alta puntuación en otra prueba. Si la deficiencia ocurre en una habilidad que es esencial para el desempeño en el trabajo, el solicitante elegido fracasará; no obstante, esta situación puede evitarse si se identifican una o más habilidades que son críticas para un empleo particular y sólo se aplica el corte en las pruebas correspondientes. Para la mayor parte de las pruebas eso suele ser preferible a conservar la puntuación real, ya que entre mayor sea la puntuación de la prueba mejor será la ejecución del individuo en el trabajo. En la mayor parte de los empleos existe una relación lineal entre la predicción de la puntuación y la ejecución en el criterio. Debe agregarse que la investigación extensa con la BTAG ha proporcionado numerosas evidencias que apoyan la existencia de dicha relación lineal (Coward y Sackett, 1990; Hartigan y Wigdor, 1989; Hawk, 1970). En esas condiciones, seleccionar a individuos sobre la base de la magnitud real de sus

[7] En el capítulo 17 estudiamos esta conocida batería en relación con el uso industrial y organizacional de las pruebas.

puntuaciones en la prueba produce un mejor desempeño en el puesto que aceptar a todos los que superan puntuaciones mínimas de corte.

USO DE LOS TESTS PARA DECISIONES DE CLASIFICACIÓN

La naturaleza de la clasificación. Las pruebas psicológicas pueden usarse con propósitos de selección, colocación o clasificación. En la *selección*, cada individuo es aceptado o rechazado; algunos ejemplos son decidir si se admite a un estudiante en la universidad, si se contrata a un solicitante o si se acepta a un recluta para un curso de entrenamiento de oficiales. Cuando la selección se hace de manera secuencial, las primeras etapas se consideran un "filtro", por lo que se reserva el término "selección" para las etapas intensivas finales. También puede emplearse el término "filtrar" para designar a cualquier proceso rápido de selección, incluso cuando no es seguido de otros procedimientos de selección.

La colocación y la clasificación difieren de la selección en que nadie es rechazado o eliminado del programa: todos los individuos son asignados a los "tratamientos" apropiados para elevar al máximo la eficacia de los resultados. En la *colocación* la asignación se basa en una sola puntuación que puede derivarse de una sola prueba, como la de aprovechamiento en matemáticas. Si se aplica una batería de pruebas puede emplearse una puntuación compuesta calculada con una ecuación de regresión simple. Algunos ejemplos incluyen la distribución de los estudiantes de primer ingreso en diferentes grupos de matemáticas de acuerdo con las puntuaciones que hayan obtenido en la prueba de rendimiento, la asignación de solicitantes a trabajos de oficina que requieren diferentes niveles de habilidad y responsabilidad y la identificación de los pacientes psiquiátricos con "mayor o menor trastorno" para tomar decisiones sobre la terapia. Es evidente que en cada una de esas decisiones sólo se emplea un criterio y la colocación es determinada por la posición del individuo en una sola escala predictora.

Por otro lado, la *clasificación* siempre comprende dos o más criterios; por ejemplo, en una situación militar la clasificación es un problema importante porque cada individuo del personal disponible tiene que ser asignado a la especialidad en la que pueda servir mejor. Las decisiones de clasificación también se requieren en la industria, como al asignar a los nuevos empleados a programas de capacitación para distintos puestos. Otros ejemplos incluyen la consejería vocacional de estudiantes al elegir carrera y área de especialización. El consejo se basa principalmente en la clasificación, ya que al cliente se le informa de sus posibilidades de tener éxito en diferentes programas académicos u ocupaciones. De igual modo, el diagnóstico clínico es un problema de clasificación, ya que el propósito principal de cada diagnóstico consiste en tomar una decisión sobre la clase más apropiada de terapia.

Aunque la colocación puede hacerse con uno o más predictores, la clasificación requiere de múltiples predictores cuya validez se determine individualmente contra cada criterio. Una batería de clasificación requiere una ecuación de regresión diferente para cada criterio. Algunas de las pruebas pueden tener pesos en todas las ecuaciones, aunque con valores distintos; otras pueden ser incluidas en sólo una o dos ecuaciones y tener un peso cero o insignificante en algunos de los criterios. De modo que la combinación de pruebas empleadas, así como los pesos específicos, difiere de acuerdo con

el criterio empleado. Un ejemplo temprano de una batería de clasificación es la que preparó la fuerza aérea estadounidense para la asignación de personal a diferentes programas de entrenamiento. El uso de esta batería, formada por pruebas de lápiz y papel y aparatos, permitió obtener calificaciones estaninas para pilotos, navegantes, bombarderos y otras especialidades de la fuerza aérea. Al encontrar una estimación de las puntuaciones de criterio individuales a partir de diferentes ecuaciones de regresión fue posible predecir, por ejemplo, si un individuo estaba mejor calificado como piloto que como navegante. Un sistema más extenso y prolongado es el Proyecto A de Selección y Clasificación del Ejército de los Estados Unidos (J. P. Campbell, 1990b).

Validez diferencial. En la evaluación de una batería de clasificación, la principal consideración es su validez diferencial comparada con criterios separados. El objeto de la batería es predecir la *diferencia* en la ejecución de cada persona en dos o más empleos, programas de capacitación u otras situaciones de criterio. Las pruebas que se eligen para integrar la batería producen coeficientes de validez muy diferentes para los criterios separados; por ejemplo, en un problema de clasificación de dos criterios la prueba ideal tendría una correlación elevada con un criterio y una correlación cero (o de preferencia negativa) con el otro. Los tests de inteligencia general son relativamente pobres para los propósitos de clasificación porque su predicción del éxito en la mayor parte de las áreas es más o menos igual de buena; por consecuencia, sus correlaciones con los criterios que deben ser diferenciados son demasiado similares. Un individuo que califique alto en esa prueba sería clasificado como exitoso en cualquier tarea, lo que vuelve imposible predecir en cuál haría mejor papel. En una batería de clasificación se necesitan algunas pruebas que sean buenas predictoras del criterio A y malas predictoras del criterio B, y otras pruebas que sean malas predictoras de A y buenas de B.

Se han elaborado procedimientos estadísticos para seleccionar pruebas que eleven al máximo la validez diferencial de una batería de clasificación (Brogden, 1946a, 1951, 1954; Horst, 1954; Mollenkopf, 1950b; Zeidner y Johnson, 1991), pero cuando el número de criterios es mayor a dos, el problema se vuelve mucho más complejo. En la práctica, se han seguido diversos procedimientos empíricos para aproximarse a las metas deseadas. J. P. Campbell (1990a, pp. 715–721) realizó un minucioso análisis de la complejidad de los problemas de clasificación.

Funciones discriminantes múltiples. Otra forma de manejar las decisiones de clasificación es la función discriminante múltiple (French, 1966), un procedimiento matemático que permite determinar qué tanto se acercan las puntuaciones del individuo en el conjunto de pruebas que caracterizan a las personas en una determinada ocupación, en algún programa de estudios, en un síndrome psiquiátrico u otra categoría. Entonces, el sujeto se asigna al grupo al que más se asemeje. Aunque la ecuación de regresión permite la predicción del grado de éxito en cada campo, la función discriminante múltiple trata a todas las personas de la misma categoría como de igual estatus. La pertenencia al grupo es el único dato de criterio que utiliza este método, lo que lo hace útil cuando no se dispone de puntuaciones de criterio y sólo puede asegurarse la pertenencia al grupo; por ejemplo, algunas pruebas se validan al aplicarlas a personas de diferentes ocupaciones, aunque no se dispone de ninguna medida de éxito laboral para individuos de cada campo.

La función discriminante también es apropiada cuando existe una relación no lineal entre el criterio y uno o más predictores; por ejemplo, en ciertos rasgos de personalidad puede haber un rango óptimo para una determinada ocupación, por lo que los individuos que tienen más o menos el rasgo en cuestión pueden estar en desventaja. Así, parece razonable esperar que los vendedores que muestren una tendencia moderadamente alta a la dominancia social tengan mayor probabilidad de éxito, la cual va disminuyendo conforme las puntuaciones se alejan de esta región en cualquier dirección. Con las funciones discriminantes tendemos a seleccionar individuos que caen dentro del rango óptimo, mientras que con la ecuación de regresión entre mayor sea la dominancia de la puntuación, más favorable se predice que será el resultado. Si la correlación entre predictor y criterio fuera negativa, es evidente que la ecuación de regresión haría predicciones más favorables para los que obtienen menores puntuaciones; pero no hay forma directa por la que una calificación intermedia reciba un crédito máximo. Aunque en muchos casos es posible llegar a la misma decisión con ambas técnicas, hay situaciones en las que las personas serían clasificadas de manera diferente por las ecuaciones de regresión y las funciones discriminantes. Para la mayor parte de los propósitos de la examinación las ecuaciones de regresión proporcionan una técnica más eficaz, pero en algunos casos la función discriminante resulta más conveniente para obtener la información requerida.

Maximizar la utilización del talento. La predicción diferencial de criterios con una batería de pruebas permite utilizar los recursos humanos de manera más completa de lo que es posible con una sola prueba general o con una puntuación compuesta de una sola ecuación de regresión. Como vimos en las tablas de Taylor-Russell y en otras partes del capítulo, la eficacia de cualquier prueba en la selección de personal para cualquier trabajo depende de la razón de selección. Al tomar decisiones de clasificación trabajamos con una razón más pequeña, lo que nos permite asignar a cada puesto a las personas mejor calificadas. Si de 100 solicitantes se necesitan 10 para cubrir dos trabajos diferentes, la razón de selección es de 10 por ciento para cada puesto cuando se emplean predictores separados para cada uno. Si se empleara un solo predictor (como un test de inteligencia general) para elegir a los solicitantes para ambos trabajos, la razón de selección sería de 20 por ciento, por lo que no podemos hacer nada mejor que aceptar a los 20 mejores solicitantes.

Incluso cuando los predictores para los dos trabajos están muy correlacionados, por lo que algunos de los solicitantes pueden calificar para ambos, hay una ganancia considerable al usar predictores separados. Esta situación se ilustra en la tabla 6.3, que muestra la media de la calificación de criterio estándar de los trabajadores seleccionados para cada uno de los empleos con una estrategia de selección (un solo predictor) y una estrategia de clasificación que incluye dos predictores distintos, cada uno validado contra su propio criterio laboral. Si los trabajadores se asignaran al azar, sin selección, la media de la calificación estándar en esta escala sería de cero. Tal sería el caso si la razón de selección fuera de 50 por ciento para cada trabajo y en consecuencia tuviera que contratarse a todos los solicitantes. Note que incluso en esas condiciones el desempeño laboral mejora con el uso de dos predictores, como se muestra en la última hilera de la tabla. Con dos predictores no correlacionados, la media del desempeño laboral sería de .31 (aproximadamente 1/3 de una desviación estándar por encima del valor por azar).

Tabla 6.3

Media estándar de la puntuación de criterio de personas colocadas en dos trabajos mediante estrategias de selección o de clasificación

Razón de selección para cada trabajo	Selección: predictor único	Clasificación: dos predictores cuya intercorrelación es				
		0	.20	.40	.60	.80
5%	.88	1.03	1.02	1.01	1.00	.96
10	.70	.87	.86	.84	.82	.79
20	.48	.68	.67	.65	.62	.59
30	.32	.55	.53	.50	.46	.43
40	.18	.42	.41	.37	.34	.29
50	.00	.31	.28	.25	.22	.17

(Adaptado de Brogden, 1951, pág. 182.)

Conforme aumenta la correlación entre los predictores, disminuye la eficacia en el trabajo de los empleados seleccionados, pero sigue siendo mejor que el azar incluso cuando la correlación es de .80. Con razones de selección menores, obviamente es posible obtener personal mejor calificado; sin embargo, como puede verse en la tabla 6.3, para cada razón de selección la ejecución media en el trabajo es mejor cuando los solicitantes se eligen mediante estrategias de clasificación que por estrategias de selección.

Un ejemplo práctico de las ventajas de las estrategias de clasificación se encuentra en el uso de las puntuaciones de áreas de aptitud en la asignación del personal a las especialidades militares en el ejército estadounidense (Maier y Fuchs, 1973). En esta investigación, cada área corresponde a un grupo de los trabajos del ejército que requiere un patrón similar de aptitudes, conocimientos e intereses. A partir de una batería de clasificación compuesta por 13 pruebas, se usan combinaciones de tres a cinco para encontrar la puntuación del individuo en cada área. La figura 6.4 muestra los resultados de una investigación de 7 500 individuos que solicitaban ser reclutados y que comparó el uso de las puntuaciones obtenidas en el área de aptitud con el uso de una prueba global de selección, el Test de Calificación de las Fuerzas Armadas (*Armed Forces Qualification Test, AFQT*). Advierta que únicamente el 56 por ciento de este grupo alcanzó o superó el percentil 50 en el AFQT, mientras que 80 por ciento alcanzó o superó la calificación estándar promedio de 100 en su mejor área de aptitud. Así, cuando los individuos se asignan a trabajos específicos de acuerdo con las aptitudes que cada puesto demanda, una gran mayoría es capaz de realizarlos tan bien o mejor que el promedio de la muestra. Esta aparente situación imposible, en la que casi cualquiera puede estar por encima del promedio, puede obtenerse al sacar provecho del hecho de que casi cualquiera destaca en *alguna* aptitud.

El mismo detalle fue ilustrado con una población muy diferente en un estudio de niños sobredotados realizado por Feldman y Bratton (1972). Con propósitos de demostración, 49 niños de dos grupos de quinto grado fueron evaluados en 19 medidas,

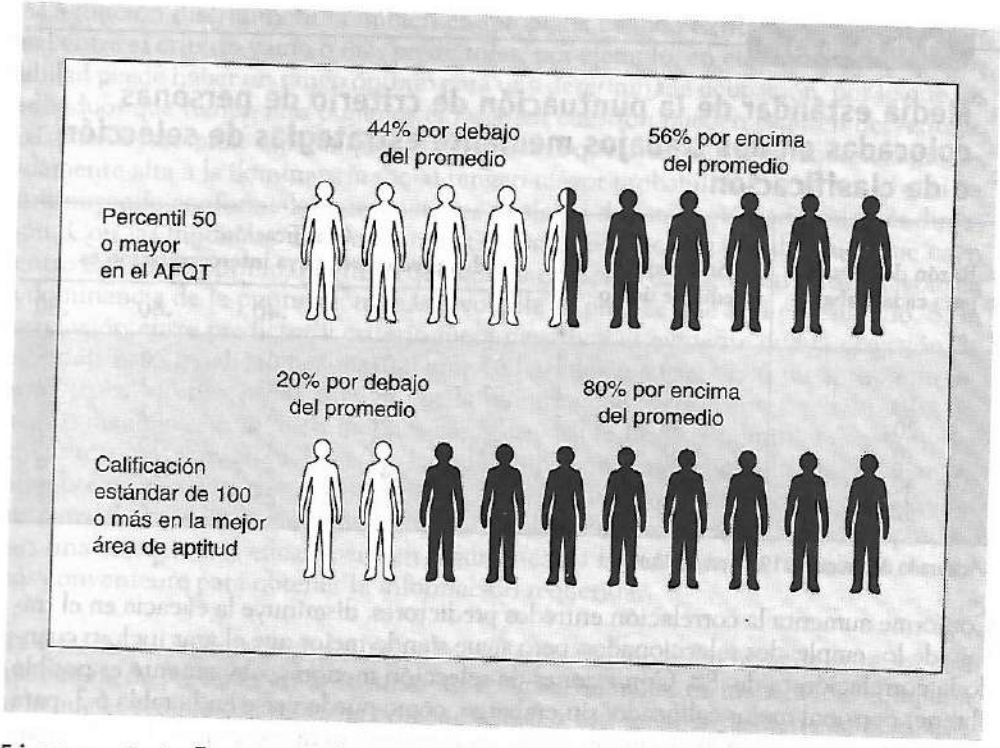


Figura 6.4. Porcentajes de puntuaciones por encima del promedio en el AFQT y en el área de mejor aptitud de la batería de clasificación del ejército en una muestra de 7 500 solicitantes de reclutamiento.

(Datos del Instituto de Investigación del Ejército de los Estados Unidos para la Conducta y Ciencias Sociales. Cortesía de J. E. Uhlaner)

todas usadas antes para seleccionar estudiantes para programas especiales de sobredotados. Entre las medidas usadas se encuentran las puntuaciones globales obtenidas en una prueba colectiva de inteligencia y en una batería de rendimiento académico, pruebas de aptitudes y de áreas académicas separadas, como lectura y aritmética, una prueba de pensamiento creativo, calificaciones en música y arte, y nominaciones de los maestros de los niños más "dotados" y de los más "creativos" en cada grupo. Cuando se identificó a los cinco niños de mayor rango seleccionados por cada criterio, se incluyó al 92 por ciento del grupo, lo que muestra una vez más que casi todos los miembros de un grupo pueden destacar cuando se emplean criterios multivariados.

ANÁLISIS ESTADÍSTICO DEL SESGO DE LA PRUEBA

El problema. Si deseamos utilizar las pruebas para predecir resultados en alguna situación futura, como el desempeño de un aspirante a la universidad o en algún trabajo, necesitamos instrumentos con alta validez predictiva del criterio particular. Este requisito suele descuidarse en el desarrollo de las llamadas "pruebas libres de cultura" (que vere-

mos en los capítulos 9 y 12). En un esfuerzo por incluir en esos instrumentos únicamente las funciones comunes a diferentes culturas o subculturas, puede elegirse un contenido que tenga poca relevancia para el criterio que se pretenda predecir. Una mejor solución consiste en elegir un contenido relevante para el criterio e investigar luego las posibles diferencias poblacionales de la efectividad de la prueba para el propósito pretendido. Los coeficientes de validez, los pesos o valores de regresión y las puntuaciones de corte pueden variar como función de las diferencias en los antecedentes de los examinados. Estas medidas deben ser por lo tanto supervisadas dentro de los subgrupos en los que hay razón para esperar dichos efectos. Las posibles diferencias en los subgrupos se reconocerán como un caso especial de la función de las variables moderadoras que estudiamos en una sección precedente. Recuerde además que la búsqueda de efectos moderadores estables y significativos ha resultado decepcionante. En esta sección, examinaremos algunas aplicaciones específicas de esa clase de análisis en poblaciones minoritarias de los Estados Unidos.

Hay que observar que es menos probable que las características predictivas de las puntuaciones de las pruebas varíen entre grupos culturales cuando la prueba es intrínsecamente adecuada para el criterio de desempeño. Si se emplea una prueba verbal para predecir la ejecución no verbal, puede encontrarse una validez fortuita en un grupo cultural debido a las asociaciones tradicionales de la experiencia dentro de esa cultura, pero la validez de esa prueba puede desaparecer en un grupo con antecedentes distintos. Por otro lado, es probable que una prueba que muestree directamente la conducta del criterio o que mida las habilidades que son requisito esencial mantenga su validez en grupos diferentes.

Desde mediados de la década de los sesenta se ha observado una rápida acumulación de investigaciones sobre posibles diferencias étnicas en el significado predictivo de las calificaciones de las pruebas.[8] La mayor parte de los estudios realizados hasta ahora son con afroamericanos, aunque algunos han incluido a otras minorías. En general, los problemas investigados se resumen bajo el encabezado de *sesgo de la prueba*. En este contexto, el término "sesgo" se emplea en su bien establecido sentido estadístico, para designar un error constante o sistemático en contraste con uno que se deba al azar. En este sentido hablamos de una muestra sesgada en contraste con una muestra aleatoria. Las principales preguntas que se han planteado en relación con el sesgo de la prueba tienen que ver con los coeficientes de validez (sesgo de la pendiente) y la relación entre las medias del grupo en la prueba y en el criterio (sesgo de intersección). Examinaremos estas cuestiones en las dos secciones siguientes.

Sesgo de la pendiente. Para facilitar la comprensión de los aspectos técnicos del sesgo de la prueba, comencemos con un diagrama de dispersión o distribución bivariada, como la que presentamos en el capítulo 4 (especialmente la figura 4.3). Para el propósito actual, el eje horizontal (X) representa las puntuaciones obtenidas en la prueba y el eje vertical (Y) las puntuaciones del criterio, como el promedio de puntuaciones en la universidad o un índice del desempeño laboral. Recordará que las marcas, que muestran la posición de cada individuo en la prueba y el criterio, indican la dirección y mag-

[8] De esta voluminosa bibliografía sólo pueden citarse unos cuantos estudios representativos. Para una revisión general del problema y un análisis de sus múltiples facetas, recomendamos N. S. Cole y Moss (1989), Hunter, Schmidt y Rauschenberger (1977), y C. R. Reynolds y Brown (1984).

nitud general de la correlación entre las dos variables. La línea de mejor ajuste que pasa por esas marcas se conoce como línea de regresión y su ecuación es la de regresión. En este ejemplo, la ecuación de regresión tendrá un solo predictor. Las ecuaciones de regresión múltiple que ya vimos tienen varios predictores, pero el principio es el mismo.

Cuando las puntuaciones de la prueba y del criterio se expresan como calificaciones estándares ($DE = 1.00$), la *pendiente* de la línea de regresión es igual al coeficiente de correlación. Es por ello que si la prueba produce un coeficiente de validez significativamente diferente en los dos grupos, la diferencia se describe como sesgo de la pendiente. Esta clase de diferencia entre grupos a menudo se conoce como "validez diferencial". Varios investigadores han empleado también el término "validez de un solo grupo" para referirse a una prueba cuyo coeficiente de validez alcanza significancia estadística en un grupo pero no en otro.

La figura 6.5 presenta ilustraciones esquemáticas de líneas de regresión para varias distribuciones bivariadas.[9] Las elipses representan la región en la que caen las marcas de cada muestra. El primer caso muestra la distribución bivariada de dos grupos con medias diferentes en el predictor, pero con líneas de regresión idénticas entre el predictor y el criterio. En este caso, no existe sesgo en la prueba porque a cualquier puntuación obtenida en ella (X) corresponde una puntuación de criterio idéntica (Y) en ambos grupos. El segundo caso ilustra el sesgo de la pendiente, con menor coeficiente de validez en el grupo minoritario.

En los estudios de validez diferencial, surge un problema común del hecho de que el número de casos en la muestra minoritaria a menudo es mucho menor que en la muestra de la mayoría. En esas condiciones, el mismo coeficiente de validez puede ser estadísticamente significativo en la muestra de la mayoría pero no en la de la minoría (llamada por eso validez de un solo grupo); por ejemplo, con 100 casos, una correlación de .27 es claramente significativa al nivel .01; con 30 casos la misma correlación queda muy lejos del mínimo requerido para alcanzar la significancia, incluso al nivel de .05. Por esta razón, el procedimiento adecuado a seguir en esos estudios de validación diferencial consiste en evaluar la diferencia entre los dos coeficientes de validez en lugar de probar por separado la significancia de cada uno (Humphreys, 1973). Con el último procedimiento sería posible "demostrar" que una prueba es válida, digamos, para blancos pero no para negros. ¡Todo lo que se necesita para lograr este propósito es un grupo lo bastante grande de blancos y lo bastante pequeño de negros!

Un sofisticado análisis estadístico de los resultados de 19 estudios publicados que informaban coeficientes de validez para muestras de empleados negros y blancos arroja serias dudas sobre las conclusiones alcanzadas en algunos de los primeros estudios (Schmidt, Berner y Hunter, 1973). Tomando en consideración la validez obtenida y el tamaño de la muestra en cada estudio, los investigadores demostraron que las discrepancias en los coeficientes de validez encontrados entre negros y blancos no difería de lo que se esperaba por azar. Esta conclusión fue corroborada en un análisis posterior más amplio que abarcaba 39 estudios (Hunter, Schmidt y Hunter, 1979). El tema de la

[9] El análisis del sesgo de la prueba que se ilustra en la figura 6.5 ha llegado a conocerse como "modelo Cleary" porque fue empleado por Cleary (1968) en un conocido estudio de la Prueba de Aptitud Académica de la Junta Universitaria sobre los resultados de estudiantes pertenecientes a minorías. Los procedimientos matemáticos se deben a Gulliksen y Wilks (1950), y Humphreys (1952) sugirió su aplicación a las comparaciones étnicas y de género. Los diagramas en la figura 6.5 se adaptaron de un estudio de M. A. Gordon (1953) conducido en la fuerza aérea estadounidense bajo la dirección de Humphrey.

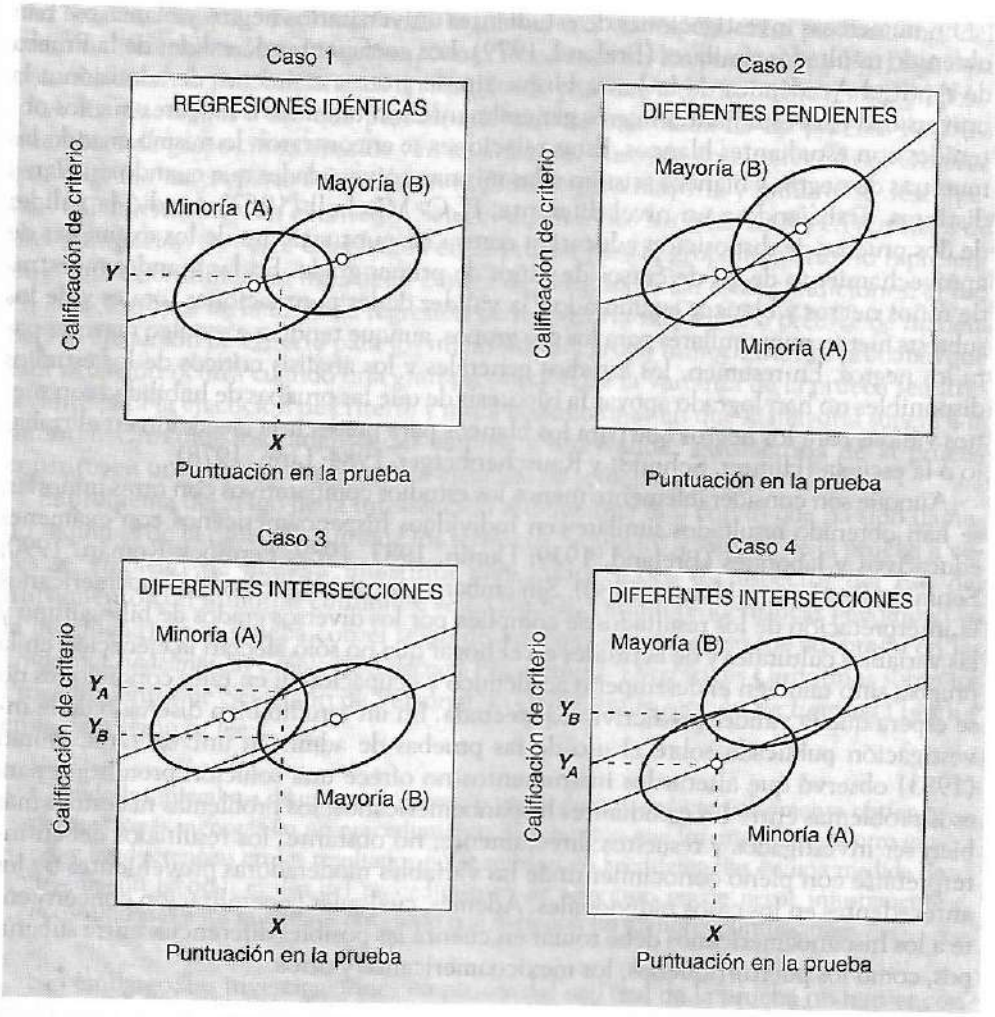


Figura 6.5. Sesgos de la pendiente e intersección en la predicción de puntuaciones de criterio. Las elipses muestran las regiones en que cayeron los miembros de cada grupo cuando sus puntuaciones se graficaron contra su ejecución del criterio.

(Los casos 1, 2 y 4 se adaptaron de M. A. Gordon, 1953, p. 3.)

validez diferencial para solicitantes de empleo pertenecientes a minorías y a grupos no minoritarios produjo una discusión continua por más de una década. Algunos investigadores observaron que, debido a las limitaciones metodológicas, los resultados sencillamente no eran concluyentes; sin embargo, es digno de mención que no se descubrió evidencia de validez diferencial en estudios bien diseñados a gran escala con muestras industriales (J. T. Campbell, Crooks, Mahoney y Rock, 1973) y con personal militar (Maier y Fuchs, 1973). En general, se demostró que era menos probable encontrar validez diferencial en los estudios metodológicamente adecuados.

En numerosas investigaciones de estudiantes universitarios negros y blancos se han obtenido resultados similares (Breland, 1979). Los coeficientes de validez de la Prueba de Aptitud Académica de la Junta Universitaria y otros exámenes de admisión a la universidad para estudiantes negros generalmente son tan altos o mayores que los obtenidos con estudiantes blancos. Estas relaciones se encontraron lo mismo cuando las muestras de negros y blancos asistían a las mismas universidades que cuando asistían a distintas. Trabajando a un nivel diferente, B. C. Mitchell (1967) estudió la validez de dos pruebas de disposición educativa contra las puntuaciones de los exámenes de aprovechamiento de fin de cursos de niños de primer grado. En las grandes muestras de niños negros y blancos examinados, la validez de las puntuaciones totales y de los subtests fueron muy similares para los dos grupos, aunque tendían a ser algo mayores para los negros. En resumen, los estudios generales y los análisis críticos de los estudios disponibles no han logrado apoyar la hipótesis de que las pruebas de habilidad son menos válidas para los negros que para los blancos para predecir la ejecución en el trabajo o la escuela (Hunter, Schmidt y Rauschenberger, 1984; Linn, 1978).

Aunque son considerablemente menos los estudios comparativos con otras minorías, se han obtenido resultados similares en individuos hispanoamericanos con exámenes educativos y laborales (Breland, 1979; Durán, 1983, 1989; Pennock-Román, 1990; Schmidt, Pearlman y Hunter, 1980). Sin embargo, en el caso de los hispanoamericanos la interpretación de los resultados se complica por los diversos grados de bilingüismo y las variables culturales y de actitudes en el hogar que no sólo afectan la ejecución en la prueba, sino también el desempeño académico y ocupacional; en tales condiciones no se espera que la validez predictiva sea afectada. En un estudio bien diseñado de la investigación publicada sobre el uso de las pruebas de admisión universitaria, Duran (1983) observó que alterar los instrumentos no ofrece una solución promisoria para esos problemas entre los estudiantes hispanoamericanos; los problemas necesitan más bien ser investigados y resueltos directamente; no obstante, los resultados deben interpretarse con pleno conocimiento de las variables moderadoras provenientes de los antecedentes en los casos individuales. Además, cualquier generalización concerniente a los hispanoamericanos debe tomar en cuenta las posibles diferencias entre subgrupos, como los puertorriqueños, los mexicoamericanos y otros.

Sesgo de intersección. Aun cuando una prueba produzca los mismos coeficientes de validez para dos grupos, es posible demostrar sesgo de intersección. La intersección de una línea de regresión se refiere al punto en el que la línea corta o se cruza con el eje. Una prueba exhibe sesgo de intersección si sistemáticamente subpredice o sobrepredice una ejecución del criterio para un grupo particular. Consideremos de nuevo el primer caso de la figura 6.5, en el que las muestras de los grupos mayoritario y minoritario muestran regresiones idénticas. En tales condiciones, no hay sesgo ni de la pendiente ni de la intersección, pues aunque los grupos difieren significativamente en las puntuaciones promedio de la prueba, muestran una diferencia correspondiente en la ejecución del criterio. Por otro lado, en el tercer caso los dos grupos tienen líneas de regresión con la misma pendiente pero diferentes intersecciones. Aquí, el grupo minoritario (A) tiene una intersección mayor que el grupo mayoritario (B); es decir, la línea de regresión del grupo minoritario interseca al eje Y en un punto más alto que la línea de regresión del grupo mayoritario. Aunque los coeficientes de validez calculados dentro de cada grupo

son iguales, cualquier puntuación de la prueba (X) corresponderá a puntuaciones del criterio diferentes en los dos grupos, como se muestra en los puntos Y_A y Y_B. La misma puntuación en la prueba tiene así un significado predictivo distinto para ambos grupos.

A los psicólogos a los que preocupa la posible inequidad de las pruebas para los miembros de grupos minoritarios ven la situación ilustrada en el tercer caso. Observe que el grupo mayoritario sobresale en la prueba, pero mayoría y minoría se desempeñan igualmente bien en el criterio. Seleccionar a todos los solicitantes en términos de una puntuación de corte establecida en la prueba para el grupo mayoritario representaría una discriminación injusta en contra de la minoría. En estas condiciones, el uso en ambos grupos de la línea de regresión de la mayoría *subpredice* o predice de manera baja la ejecución de criterio para los miembros del grupo minoritario. Es probable que esta situación ocurra cuando una gran proporción de la varianza de la prueba sea irrelevante para la ejecución de criterio y mida las funciones en que la mayoría supera a la minoría. Un cuidadoso análisis del empleo y una validez satisfactoria de la prueba constituyen una protección contra la elección de semejante prueba.

El problema del sesgo de la intersección se relaciona más con lo que ha sido llamado "equidad de la prueba". Aunque los términos "equidad" y "sesgo" de la prueba a veces se emplean de manera indistinta para cubrir todos los aspectos del uso del instrumento con minorías culturales, se ha hecho costumbre identificar la equidad (o la inequidad) de la prueba con el sesgo de intersección. Este uso fue adoptado en las *Uniform Guidelines on Employee Selection Procedures* ("Directrices Uniformes sobre los Procedimientos de Selección de Personal") (1978). En la sección de *Equidad* (14B), y en la definición se lee:

> Cuando los miembros de una raza, género o grupo étnico, característicamente obtienen puntuaciones menores en un procedimiento de selección que los miembros de otro grupo, y las diferencias en los resultados no se reflejan en las diferencias en una medida de desempeño laboral, el uso del procedimiento de selección puede negar injustamente oportunidades a los miembros del grupo que obtienen las menores puntuaciones.

Sin embargo, las investigaciones empíricas del uso real de la prueba no han encontrado sesgo de intersección significativo o, más a menudo, han hallado una ligera tendencia en dirección *opuesta*, como se ilustra en el cuarto caso de la figura 6.5: aquí, el grupo mayoritario (B) tiene una mayor intersección que el grupo minoritario (A). En estas condiciones, el uso de la línea de regresión y de la puntuación de corte de la mayoría para ambos grupos sobrepredice o predice de manera alta la ejecución del criterio de los miembros del grupo minoritario y, por ende, discrimina a su *favor*. Tales resultados se han obtenido en la predicción de calificaciones universitarias (Breland, 1979; Duran, 1983; Zeidner, 1987), calificaciones en escuelas de leyes (Linn, 1975), desempeño en programas de entrenamiento del ejército y la fuerza aérea (M. A. Gordon, 1953; Maier y Fuchs, 1973; C. W. Shore y Marion, 1972) y una amplia variedad de criterios industriales (véase Hunter *et al.*, 1984 para una revisión).

Se ha demostrado matemáticamente que el cuarto caso (figura 6.5) ocurre si dos grupos difieren en una o más *variables adicionales* que correlacionan en forma positiva con la prueba y el criterio (Linn y Werts, 1971; Reilly, 1973). La pequeña sobrepredicción es

un método estadístico para considerar un predictor tras otro. Conforme se agregan más predictores a la batería, disminuye la sobrepredicción, un hecho que ha sido verificado de modo empírico con varias poblaciones, que van de estudiantes de escuelas de leyes y empleados de oficina a pequeños del jardín de infantes (véase Hunter *et al.*, 1984).

Es interesante notar que se han obtenido los mismos resultados al hacer comparaciones entre grupos clasificados de acuerdo con el nivel educativo o socioeconómico. La Batería de Clasificación del Ejército tendía a sobrepredecir la ejecución de individuos que habían abandonado la secundaria y a subpredecir el desempeño de graduados universitarios en los programas de entrenamiento para las especialidades ocupacionales del ejército (Maier, 1972). De igual modo, diversas pruebas de aptitud académica tendían a subpredecir las calificaciones universitarias de estudiantes cuyos padres eran profesionales y a sobrepredecir las calificaciones de estudiantes de menores niveles ocupacionales (Hewer, 1965). En todos estos estudios, las comparaciones de los grupos con mayor y menor calificación reveló no encontrar diferencia significativa en la intersección o un ligero sesgo a favor de los grupos con menor calificación en la prueba.

Modelos de decisión para el uso justo de los tests. Gradualmente empezó a cambiar el interés de la investigación en la evaluación del sesgo de las pruebas al diseño de estrategias de selección para su uso justo con minorías culturales. Si una estrategia de selección sigue el modelo de regresión (véase Cleary) ilustrado en la figura 6.5, los individuos serán elegidos (para su ingreso a la universidad, al empleo, etc.) únicamente sobre la base de sus calificaciones predictivas de criterio. Esta estrategia elevará al máximo la ejecución de criterio global, sin considerar otras metas del proceso de selección. De acuerdo con esta estrategia, un uso justo de las pruebas en la selección es aquel que se basa en la mejor estimación de la ejecución del criterio para cada individuo.

Se han propuesto muchos otros modelos de decisión que tienen el efecto de seleccionar mayores proporciones del grupo con menor puntuación. Este efecto está de acuerdo con la meta que en general se conoce como "acción afirmativa" o la reducción del "impacto adverso" en el proceso de selección. Cuando se introdujeron, los modelos alternativos parecían seguir procedimientos muy diferentes a los del modelo de regresión;[10] sin embargo, posteriormente se demostró que todos podían expresarse como variantes de un modelo más amplio (Darlington, 1971; Gross y Su, 1975; Petersen y Novick, 1976). Las diferencias entre los diversos modelos pueden explicarse en términos de los juicios de valor implícitos en cada uno. La función de los valores en las estrategias de toma de decisiones se ilustró al inicio del capítulo (véase la figura 6.2). Recuerde que la asignación de una *utilidad* relativa a cada resultado requiere de un juicio sobre el grado del carácter favorable o desfavorable de dicho resultado. Esos juicios, junto con la probabilidad de cada resultado, se utilizan al calcular la utilidad total esperada (*UE*) de la estrategia de decisión.

[10] La bibliografía sobre los diversos modelos de decisión para el uso justo de las pruebas es abundante, y gran parte de ella de naturaleza técnica. Para resúmenes sencillos de los rasgos distintivos y las implicaciones de los diferentes modelos, véase Bond (1981), Dunnette y Borman (1979, pp. 497–500), Gross y Su (1975, pp. 350–351), y C. R. Reynolds (1982). Pueden encontrarse explicaciones más completas en Hunter y Schmidt (1976) y Hunter *et al.* (1977).

El análisis teórico de la decisión del uso justo de la prueba demostró que los modelos propuestos difieren en su definición de equidad o justicia en la medida en que implícitamente asignan valores diferentes a la aceptación o al rechazo de éxitos y fracasos potenciales dentro de poblaciones minoritarias y no minoritarias. El modelo de utilidad esperada hace explícitos los valores sociales implícitos y exige una declaración abierta de las utilidades, la que no puede alcanzarse por medios estadísticos sino que requiere una discusión abierta y aproximaciones sucesivas al equilibrio de metas en conflicto (N. S. Cole y Moss, 1989; Darlington, 1976; Messick, 1989). Entre las metas por reconciliar están las de proporcionar iguales oportunidades a todos los individuos, elevar al máximo la tasa de éxito y la productividad, incrementar la mezcla demográfica y la representatividad de la fuerza de trabajo (al menos en ciertas ocupaciones) y extender el tratamiento preferencial a grupos desfavorecidos por inequidades anteriores.

Finalmente, debe subrayarse que los ajustes estadísticos a las calificaciones de la prueba, puntuaciones de corte y fórmulas de predicción son poco prometedores como medio para corregir las inequidades sociales. Es posible que el uso de manipulaciones estadísticas que encubran las diferencias mediante normas separadas de subgrupos dañe a los individuos[11] al asignarlos a trabajos o programas educativos para los que carecen de las habilidades o el conocimiento requeridos. Es probable que el resultado sea un pobre desempeño, lo que no sólo afecta adversamente al individuo en su autoconcepto y actitud hacia el trabajo, sino que también puede contribuir a la perpetuación de un estereotipo de que las personas de alguna categoría cultural o étnica se desempeñan mal. Otras aproximaciones citadas antes proponen soluciones más constructivas. Una de ellas la ilustran los exámenes de aptitudes múltiples y las estrategias de clasificación que permiten la utilización plena de los diversos patrones de aptitud favorecidos por diversos grupos culturales. Una consideración más amplia de los rasgos relevantes de personalidad, motivación y de actitudes también contribuye a la predicción de la ejecución laboral o educativa. Otro planteamiento es el de tratamientos adaptados, como los programas de capacitación individualizados. Para elevar al máximo su adecuación a las características individuales, es esencial que las pruebas revelen tan plena y exactamente como sea posible el nivel actual de desarrollo de la persona en las habilidades requeridas. Los modelos generales de toma de decisiones proporcionan un marco de referencia para combinar una diversidad de métodos y sistemas de valores al evaluar la efectividad global de cada solución.

[11] Véase, por ejemplo, D. C. Brown (1994), L. S. Gottfredson (1994) y Sackett y Wilk (1994).

Capítulo Siete

Análisis de reactivos

La familiaridad con las bases conceptuales y las técnicas del análisis de ítemes o reactivos, así como el conocimiento de las otras fases del desarrollo de las pruebas, puede ayudar a los usuarios a evaluar los instrumentos publicados. El análisis de reactivos es también particularmente importante en la elaboración de pruebas locales de tipo informal, como las preguntas y exámenes que preparan los maestros para aplicar en el aula. Algunas de las directrices generales para la redacción de buenos reactivos, así como las técnicas estadísticas más sencillas de análisis, pueden mejorar los exámenes escolares y su uso ser valioso incluso con grupos pequeños.

Los reactivos pueden analizarse de modo cualitativo, en términos de su forma y contenido, y cuantitativo, en función de sus propiedades estadísticas. El análisis cualitativo incluye la consideración de la validez de contenido que revisamos en el capítulo 5, así como la evaluación de los reactivos en términos de los procedimientos adecuados de redacción. El análisis cuantitativo se refiere principalmente a la medición de la dificultad y discriminación del reactivo. La validez y la confiabilidad de cualquier prueba dependen en última instancia de las características de sus reactivos, por lo que pueden incorporarse en la prueba desde el inicio mediante el análisis de reactivos. Las pruebas pueden mejorarse mediante la selección, sustitución o revisión de los reactivos.

El análisis de reactivos permite acortar una prueba y al mismo tiempo incrementar su validez y confiabilidad. En igualdad de circunstancias, una prueba más larga es más válida y confiable que una más corta. Al presentar la fórmula de Spearman-Brown en el capítulo 4 revisamos el efecto de alargar o acortar una prueba sobre el coeficiente de confiabilidad. Esos cambios estimados en la confiabilidad ocurren cuando los reactivos eliminados son equivalentes a los que se aceptan o retienen, o bien cuando se

agregan nuevos reactivos equivalentes. Al eliminar o agregar reactivos de validez equivalente se producen cambios similares en la validez. Todos esos cambios estimados en la confiabilidad o validez se refieren al alargamiento o acortamiento de las pruebas realizado mediante una selección *aleatoria* de los reactivos, sin analizarlos; no obstante, cuando una prueba se acorta eliminando los reactivos menos válidos y confiables, la versión abreviada es más válida y confiable que el instrumento original.

DIFICULTAD DE LOS REACTIVOS

Porcentaje de aprobación. Para la mayor parte de los propósitos de examinación, la dificultad de un reactivo se define en términos del porcentaje (o proporción) de personas que lo contestan correctamente. Entre más sencillo sea el reactivo más elevado será el porcentaje. Una palabra que define correctamente el 70 por ciento de la muestra de estandarización ($p = .70$) se considera más fácil que una que se define apenas por el 15 por ciento ($p = .15$). Se acostumbra disponer los reactivos por orden de dificultad, de modo que el examinado empiece con los más sencillos. Este ordenamiento da al examinado confianza al aproximarse a la prueba y hace menos probable que pierda demasiado tiempo en los reactivos que están por encima de su habilidad y descuide los más sencillos que puede contestar correctamente.

En el proceso de elaboración de la prueba, la elección de los reactivos con un nivel adecuado de dificultad constituye una razón importante para medir la dificultad de los reactivos. La mayor parte de las pruebas estandarizadas de habilidad están diseñadas para evaluar tan exactamente como sea posible el nivel concreto alcanzado por el individuo; por ello, se considera a los reactivos que nadie aprueba, o que todos aciertan, como una especie de exceso de equipaje, puesto que ninguno proporciona información sobre las diferencias individuales. Dado que esos reactivos no afectan la variabilidad de las puntuaciones, en nada contribuyen a la validez de la prueba. Entre más se acerque a 1.00 o a 0 el grado de dificultad del reactivo, menor es la información diferencial que aporta sobre los examinados; por el contrario, la diferenciación que puede hacer aumenta conforme su dificultad se acerca a .50. Suponga que de 100 personas 50 aprueban un reactivo y 50 lo fallan ($p = .50$). Este reactivo nos permite diferenciar entre cada uno de los que lo aprobaron y cada uno de los que lo fallaron. Tenemos así 50×50 o 2 500 comparaciones apareadas o piezas de información diferencial. Un reactivo aprobado por el 70 por ciento ofrece 70×30 o 2 100 piezas de información; uno aprobado por 90 por ciento brinda 90×10 o 900; mientras que uno aprobado por todos ofrece 100×0 o 0. La misma relación puede sostenerse en el caso de los reactivos más difíciles, que son aprobados por menos del 50 por ciento.

Parecería entonces que para obtener una diferenciación máxima debemos elegir todos los reactivos con un grado de dificultad de .50; pero la decisión se complica por el hecho de que en la prueba los reactivos tienden a estar correlacionados, y entre más homogéneo sea el instrumento, mayores serán estas correlaciones. En el caso extremo, si todos los reactivos estuvieran perfectamente correlacionados y todos correspondieran al grado de dificultad de .50, las mismas 50 personas del grupo de 100 aprobarían cada reactivo. En consecuencia, una mitad de los examinados obtendría calificaciones perfectas y la otra de cero. Debido a las correlaciones entre los reactivos es mejor seleccionar a los que muestran una dispersión moderada del grado de dificultad, pero cuyo

promedio de dificultad es de .50. Más aún, entre más grandes sean las correlaciones de los reactivos (o sus correlaciones con la puntuación total), mayor debe ser la dispersión de la dificultad de los reactivos.

Otro elemento por considerar al elegir los reactivos de dificultad apropiada tiene que ver con la probabilidad de adivinación en los reactivos de opción múltiple. Para tener en consideración el hecho de que cierta proporción de examinados adivinará la opción correcta, se establece una proporción deseada de respuestas correctas mayor de lo que sería el caso para un reactivo de respuesta libre; por ejemplo, para un reactivo con cinco opciones, la proporción promedio de respuestas correctas será aproximadamente de .69 (Lord, 1952).

Escalas de intervalo. El porcentaje de quienes aciertan un reactivo indica su dificultad en términos de una escala ordinal, es decir, señala correctamente el orden de clasificación o de dificultad relativa de los reactivos; por ejemplo, si los reactivos uno, dos y tres son contestados correctamente por 30, 20 y 10 por ciento de los casos, respectivamente, podemos concluir que el primer reactivo es el más fácil y el tercero el más difícil. Pero no podemos inferir que la diferencia en dificultad entre los reactivos uno y dos es igual a la que existe entre los reactivos dos y tres. Iguales porcentajes de diferencias corresponden a iguales diferencias de dificultad únicamente en una distribución rectangular en la que los casos se distribuyan uniformemente. Este problema es similar al que se enfrenta en relación con las puntuaciones percentilares, que también se basan en porcentajes de casos. Recordará que en el capítulo 3 vimos que las puntuaciones percentilares no representan unidades iguales, sino que difieren en magnitud del centro a los extremos de la distribución (figura 3.4).

Si suponemos que el rasgo o atributo medido por cualquier reactivo tiene una distribución normal, el grado de dificultad del reactivo puede expresarse en términos de una escala de intervalo en unidades iguales luego de consultar la tabla de la curva normal de frecuencias; por ejemplo, en el capítulo 3 ya vimos que en una distribución normal aproximadamente 34 por ciento de los casos se localiza entre la media y una distancia de 1σ en cualquier dirección (figura 3.3). Con esta información podemos examinar la figura 7.1, que muestra el grado de dificultad de un reactivo aprobado por el 84 por ciento de los casos. Como las personas de la parte superior de la distribución son las que aprueban y las de la parte inferior las que fallan, este porcentaje incluye a la mitad superior (50 por ciento) más el 34 por ciento de los casos de la mitad inferior ($50 + 34 = 84$). De ahí que, como se muestra en la figura 7.1, el reactivo cae a 1σ *por debajo* de la media. El reactivo aprobado por el 16 por ciento de los casos se ubica en 1σ *por encima* de la media, ya que ahí se encuentra el 16 por ciento de los casos ($50 - 34 = 16$). El reactivo aprobado exactamente por la mitad de los casos cae en la media, por lo que en esta escala tendría un valor de 0. Los reactivos de mayor dificultad tienen valores mayores positivos y los más sencillos valores menores negativos. El índice de dificultad correspondiente a cualquier porcentaje de aprobación puede encontrarse en la tabla de la curva normal de frecuencias de cualquier libro de estadística.

Escalamiento absoluto de Thurstone. Los índices de dificultad de los reactivos que se expresan como porcentajes o unidades de la curva normal se limitan al rango de habilidad cubierto por la muestra de la que fueron obtenidos; sin embargo, en algu-

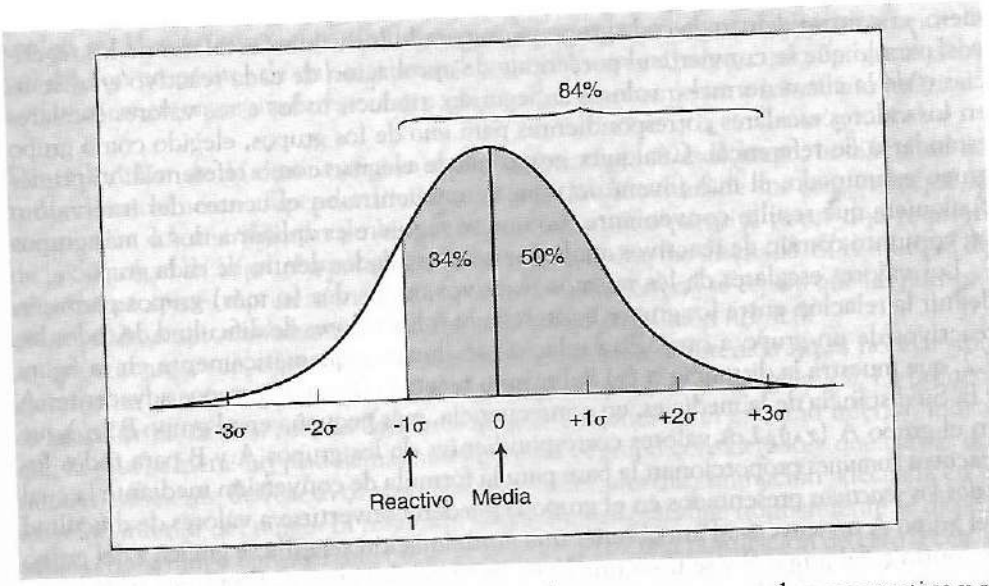


Figura 7.1. Relación entre el porcentaje de personas que aprueban un reactivo y el grado de dificultad del reactivo en Unidades de la Curva Normal.

nos casos es necesario contar con una medida de la dificultad del reactivo que pueda aplicarse a muestras cuyo nivel de habilidad varíe; por ejemplo, en las pruebas de rendimiento académico resulta ventajoso poder comparar la puntuación de un niño en varios grados escolares sucesivos con una escala uniforme. Pero obviamente no sería práctico escalar los reactivos apropiados para todos los grados aplicándolos a un solo grupo, ya que algunos resultarían demasiado difíciles y otros demasiado sencillos para casi cualquiera del grupo.

Otro ejemplo se encuentra en los programas de examinación a gran escala, como los de ingreso a la universidad, que requieren la aplicación de muchas formas equivalentes en momentos distintos. En el capítulo 3 consideramos la forma en que este problema influye en la interpretación de las puntuaciones totales de los instrumentos como la Prueba de Evaluación Académica. La solución descrita en ese caso consistió en usar un grupo estándar de referencia para definir las unidades y el origen de la escala y convertir luego a esta escala todas las puntuaciones subsecuentes. Esta conversión requiere la inclusión de un conjunto de reactivos ancla o de enlace en las pruebas aplicadas a cualquier par de grupos. Los reactivos constituyen una miniprueba, ya que son representativos de la prueba completa en contenido y forma. Diferentes pares de grupos pueden compartir un conjunto distinto de reactivos de enlace. Cada nueva forma está ligada a una o dos formas anteriores, que a su vez se vinculan a otras mediante una cadena de minipruebas que retroceden hasta el grupo original de referencia.

Es posible emplear el mismo método general para medir la dificultad de cada reactivo en una escala uniforme aplicable a cualquier número de grupos entrelazados. El procedimiento estadístico, conocido como escalamiento absoluto, fue elaborado por Thurstone (1925, 1947a), y se le ha utilizado mucho en la elaboración de pruebas (por ejemplo, Donlon, 1984). En esencia, el procedimiento consta de dos pasos: pri-

mero, encontrar dentro de cada grupo por separado los valores escalares de los reactivos, para lo que se convierte el porcentaje de aprobación de cada reactivo en distancias σ de la curva normal o valores *z*. Segundo, traducir todos estos valores escalares en los valores escalares correspondientes para *uno* de los grupos, elegido como grupo estándar o de referencia. Cualquier grupo puede elegirse como referencia: el primer grupo examinado, el más joven, uno que se encuentre en el centro del intervalo o cualquiera que resulte conveniente. Lo que se requiere es aplicar a dos o más grupos un conjunto común de reactivos ancla que sean escalados dentro de cada grupo.

Los valores escalares de los mismos reactivos en los dos (o más) grupos permiten definir la relación entre los grupos e intercambiar los valores de dificultad de todos los reactivos de un grupo a otro. Esta relación se ilustra esquemáticamente en la figura 7.2, que muestra la distancia σ (z) del mismo reactivo (i) en dos grupos adyacentes A y B. Su distancia de la media es, en consecuencia, más pequeña en el grupo B (z_B) que en el grupo A (z_A). Los valores correspondientes de los grupos A y B para todos los *reactivos comunes* proporcionan la base para la fórmula de conversión mediante la cual *todos los reactivos* presentados en el grupo B pueden convertirse a valores de dificultad del grupo A o viceversa. Para obtener una aproximación sencilla se grafica *z* del grupo A contra *z* del grupo B y se traza una línea recta por los puntos. Esta línea sirve luego para encontrar los valores z_A para todos los reactivos aplicados al grupo B.

El mismo procedimiento de conversión puede extenderse a cualquier número de grupos, trabajando con pares de grupos adyacentes que se traslapan; por ejemplo, en una prueba diseñada para aplicarse del primero al octavo grado, los valores escalares del octavo grado pueden trasladarse a la escala del séptimo grado, la escala de éste a la del sexto, y así sucesivamente hasta el primer grado. Los grupos de grados adyacentes

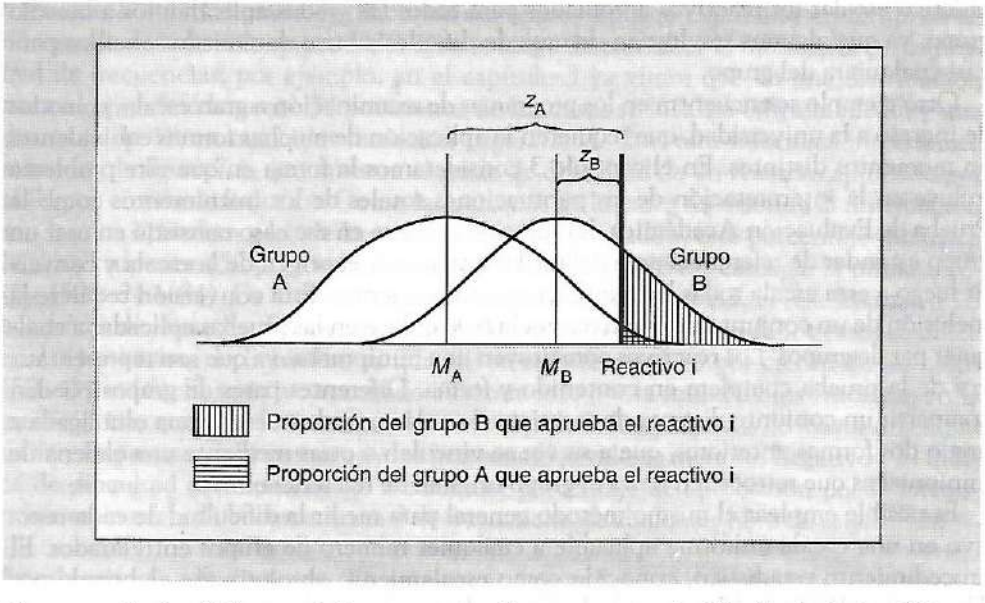


Figura 7.2. Valores *z* de la curva normal que muestran la dificultad relativa del mismo reactivo en los grupos A y B.

suelen ser lo suficientemente similares para compartir un segmento importante de la prueba con propósitos de enlazamiento o conexión; sin embargo, cualquier grado debe compartir segmentos similares con los grados inmediato superior e inmediato inferior.

Distribución de las puntuaciones de la prueba. Desde luego, la dificultad de todo el instrumento depende directamente de la dificultad de los reactivos que la componen. La distribución de las puntuaciones totales permite hacer una revisión de la dificultad de la prueba para la población para la que fue diseñada. Si la muestra de estandarización es representativa de la población, en general se espera que las puntuaciones caigan aproximadamente en una curva de distribución normal.

Sin embargo, supongamos que la curva de distribución obtenida no es normal sino claramente asimétrica, como se observa en las partes A y B de la figura 7.3. La primera de esas distribuciones, con un conjunto de puntuaciones en el extremo inferior, indica que la prueba tiene un piso demasiado alto para el grupo considerado y que carece del número suficiente de reactivos sencillos para hacer una discriminación adecuada en el extremo inferior del rango. El resultado es que las personas que normalmente se dispersarían en un rango considerable obtienen en la prueba una puntuación de cero o cercana a cero, por lo que se encuentra un pico en el extremo inferior de la escala. El esquema de la figura 7.4 ilustra este conjunto artificial de puntuaciones, en la que un grupo normalmente distribuido produce una distribución asimétrica en una prueba particular. La asimetría opuesta se observa en la parte B de la figura 7.3, en la que las puntuaciones se agrupan en el extremo superior, lo que indica un techo insuficiente para la prueba. La aplicación de una prueba diseñada para la población general a muestras seleccionadas de estudiantes universitarios por lo general producirá una distribución asimétrica, en la que un número de estudiantes obtiene puntuaciones casi perfectas. Con dicha prueba es imposible medir las diferencias individuales entre los estudiantes más capaces del grupo. De haberse incluido en las pruebas reactivos más difíciles, es indudable que algunos individuos habrían obtenido puntuaciones mayores a lo que la prueba actual permite.

Cuando la muestra de estandarización produce una distribución marcadamente anormal en la prueba, su grado de dificultad suele modificarse hasta que se acerca a una curva normal. Dependiendo del tipo de desviación de la normalidad que se en-

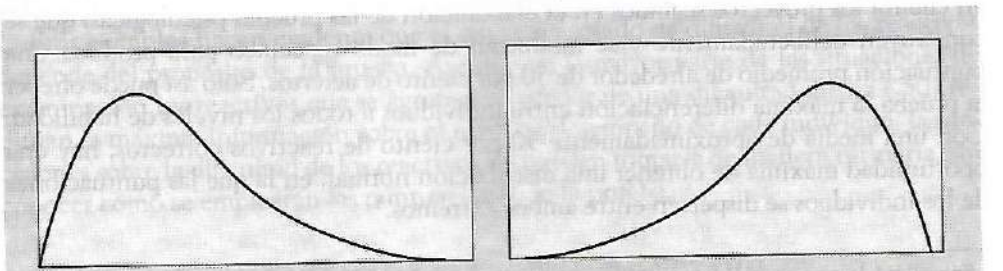


Figura 7.3. Curvas de distribución asimétricas.

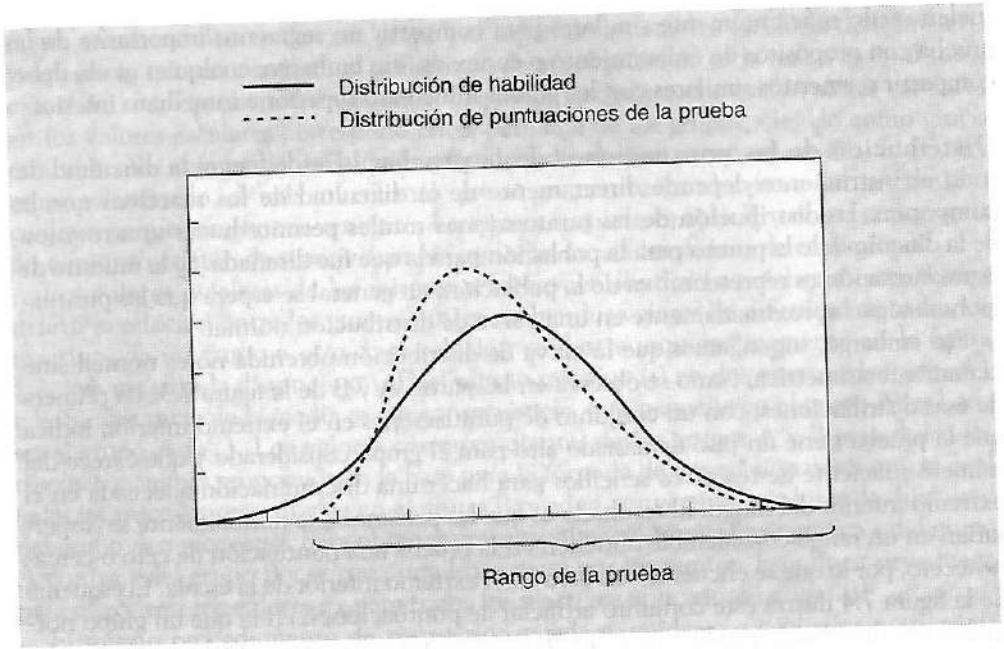


Figura 7.4. Asimetría resultante de un piso insuficiente de la prueba.

cuentre, pueden agregarse reactivos más sencillos o más difíciles, eliminarse o modificarse otros, cambiarlos de posición o revisar los pesos asignados a ciertas respuestas. Estos ajustes continúan hasta que la distribución se hace al menos aproximadamente normal. En estas condiciones, la puntuación más probable, obtenida por el mayor número de personas, suele corresponder a alrededor del 50 por ciento de reactivos correctos. Para quien no esté familiarizado con los métodos de elaboración de pruebas psicológicas, una puntuación de 50 por ciento puede parecer escandalosamente baja. En ocasiones se objeta, sobre esa base, que el examinador ha establecido un criterio demasiado bajo de aprobación o bien se infiere que el grupo examinado es particularmente malo. Por supuesto, ambas conclusiones carecen de sentido cuando se tienen en cuenta los proyectos seguidos en la elaboración de las pruebas psicológicas, que se construyen deliberadamente y se modifican de modo específico para producir una puntuación promedio de alrededor de 50 por ciento de aciertos. Sólo así puede ofrecer la prueba la máxima diferenciación entre individuos a todos los niveles de habilidad. Con una media de aproximadamente 50 por ciento de reactivos correctos, hay una oportunidad máxima de obtener una distribución normal, en la que las puntuaciones de los individuos se dispersen entre ambos extremos.[1]

[1] En realidad, la curva normal permite hacer una discriminación más fina en los extremos que en el centro de la escala. Una discriminación igual en todos los puntos de la escala requeriría una distribución rectangular; sin embargo, si después se pretende realizar análisis estadísticos de las puntuaciones la curva normal tiene una ventaja, ya que muchas técnicas estadísticas suponen una normalidad aproximada de la distribución. Por ésta y otras razones es probable que durante algún tiempo la mayor parte de las pruebas diseñadas para uso general sigan el patrón de la curva normal.

Relación entre la dificultad del reactivo y los propósitos del examen. En general, las pruebas psicológicas estandarizadas se diseñan para permitir la máxima diferenciación entre individuos de todos los niveles. Hasta ahora, nuestra revisión de la dificultad de los reactivos se ha dirigido a esta clase de instrumento; sin embargo, al elaborar pruebas para propósitos especiales, tanto la elección del grado adecuado de dificultad como la forma óptima de la distribución de las puntuaciones dependen de la discriminación buscada. En consecuencia, las pruebas diseñadas con propósitos de elección deberían emplear reactivos cuyos valores de dificultad se acerquen a la razón de selección deseada; por ejemplo, para seleccionar al 20 por ciento superior de los casos, los mejores reactivos son los que se agrupan alrededor de una *p* de .20 (o algo mayor para tener en cuenta la adivinación). Como en una prueba de elección no se requiere diferenciar *dentro* de los grupos aceptados o rechazados, el uso más efectivo del tiempo de examinación se obtiene cuando los reactivos se agrupan cerca del punto crítico de corte. De lo anterior se deriva, por ejemplo, que si quiere usarse una prueba para clasificar a los solicitantes de becas de una población universitaria, la dificultad de los reactivos debe ser considerablemente mayor que el promedio para esa población. De modo similar, si debe seleccionarse a personas de lento aprendizaje para un programa de remedio, los reactivos más deseables serían los que son más sencillos que el promedio.

Las pruebas de habilidad ofrecen otro ejemplo de la elección de los grados de dificultad. Recordará (capítulo 3) que las pruebas de habilidad a menudo se asocian con la examinación referida al dominio. Si la prueba pretende averiguar si un individuo ha dominado los elementos básicos de una habilidad o si ha adquirido el conocimiento requerido para avanzar en un programa de aprendizaje, los reactivos deberían encontrarse a un nivel *p* de .80 o .90. En esas condiciones, puede esperarse que la mayoría de los examinados complete correctamente casi todos los reactivos. Así, los reactivos muy sencillos (incluso los que se aprueban en todos los casos), que las pruebas estandarizadas comunes eliminan como no discriminativos, son los mismos que se incluirían en una prueba de destreza. De modo similar, una preevaluación realizada antes de una unidad de aprendizaje para determinar si alguno de los estudiantes adquirió ya la habilidad que pretende enseñarse, producirá porcentajes muy bajos de aprobación para cada reactivo. En este caso, no se deberían eliminar los reactivos con valores *p* muy bajos o incluso de cero, ya que revelan lo que queda por ser aprendido.

Esos ejemplos hacen evidente que el grado apropiado de dificultad de los reactivos depende del propósito de la prueba. Aunque en la mayor parte de las situaciones de examinación los reactivos que se agrupan alrededor de una dificultad media (.50) reditúan la máxima información sobre el nivel de desempeño de cada individuo, las decisiones sobre la dificultad de los reactivos no pueden tomarse de manera rutinaria, sin conocer cómo se emplearán las puntuaciones de la prueba.

DISCRIMINACIÓN DEL REACTIVO

Elección del criterio. La capacidad discriminativa del reactivo se refiere al grado en que éste diferencia correctamente entre los examinados en relación con la conducta que la prueba pretende medir. Cuando debe evaluarse toda la prueba mediante la va-

lidación relacionada con el criterio, es posible evaluar y seleccionar los reactivos sobre la base de su relación con el mismo criterio externo. Este procedimiento se ha empleado en el desarrollo de algunos tests de personalidad y de interés que estudiaremos en los capítulos 13 y 14. También es el método que suele seguirse al elegir los reactivos que han de incluirse en los inventarios biográficos, que por lo general cubren una colección heterogénea de antecedentes del individuo. En estos instrumentos no existe una base previa para calificar una respuesta como correcta o incorrecta o para asignar un peso de calificación, salvo por comparación con la posición del criterio de quienes dan esa respuesta. De un banco inicial de reactivos, se conservan los que mejor distinguen entre las personas clasificadas en diferentes categorías del criterio, como ocupaciones o síndromes psiquiátricos. Con frecuencia, los grupos de criterio constan de éxitos y fracasos en un curso académico, un programa de entrenamiento o una clase de trabajo.

En las pruebas de habilidad o destreza referidas al dominio que explicamos en el capítulo 3, los reactivos pueden evaluarse al comparar el desempeño de individuos que varían en la cantidad de educación en las funciones relevantes (Panell y Laabs, 1979; L. A. Shepard, 1984). Por lo general se compara la proporción de personas que dan la respuesta correcta en un grupo antes y después de la instrucción. Como esas pruebas se utilizan para determinar si los individuos han alcanzado un nivel específico de destreza, las diferencias individuales en la ejecución global en una sola ocasión disminuyen al mínimo, por lo que el análisis interno del reactivo no será significativo y se hará necesario contar con un criterio externo, como la cantidad de educación relevante.

En otras pruebas de desempeño, como es el caso de muchos tests de aptitud, suele investigarse la capacidad discriminativa del reactivo en contraste con la puntuación total en el instrumento mismo.[2] En las pruebas de aprovechamiento académico, generalmente no se dispone de un criterio externo, mientras que en los tests de aptitud, el énfasis en la validación del constructo hace de la puntuación total un criterio apropiado para la selección de reactivos. En las etapas iniciales de la elaboración del test, la puntuación total constituye la primera aproximación a una medida de la habilidad, atributo o constructo investigado.

Examinemos además las implicaciones de elegir los reactivos sobre la base de un criterio externo y de la puntuación total de la prueba. El primero tiende a elevar al máximo la validez del instrumento con respecto al criterio externo, mientras que el segundo aumenta la consistencia interna u homogeneidad de la prueba. En algunas condiciones pueden llevar a resultados opuestos, y los reactivos que se eligen sobre la base de la validez externa son los mismos que se rechazan en función de la consistencia interna. Supongamos que la forma preliminar de una prueba de aptitud académica consta de 100 reactivos de aritmética y 50 de vocabulario. Para seleccionar los reactivos del banco inicial por el método de consistencia interna, se calcula algún índice de acuerdo entre la ejecución en cada reactivo y la puntuación total en los 150 reactivos. Es evidente que dicho índice tenderá a ser mayor para los reactivos de aritmética que para los de vocabulario, ya que la puntuación total se basa en el doble de reactivos de aritmética.

[2] Las correlaciones entre reactivo y prueba estarán algo infladas por el común específico y la varianza de error en el reactivo y la prueba de que es parte. Se dispone de fórmulas para corregir este efecto de parte-todo (Guilford y Fruchter, 1978, pp. 165–167).

Si se deseara conservar en la forma final de la prueba a los 75 "mejores" reactivos, es probable que en su mayor parte resulten ser problemas aritméticos; sin embargo, en términos del criterio externo de aprovechamiento académico, los reactivos de vocabulario podrían ser predictores más válidos que los de aritmética, en cuyo caso la validez de la prueba habrá disminuido en lugar de aumentar gracias al análisis de reactivos.

La práctica de rechazar los reactivos que tienen baja correlación con la puntuación total permite purificar u homogeneizar la prueba, ya que al aplicarla se conservan los reactivos con alto promedio de intercorrelaciones. Este método de selección de reactivos sólo incrementará la validez de la prueba cuando el banco original de reactivos mida un solo atributo y éste se encuentre presente en el criterio o constructo que se pretende evaluar; no obstante, algunas pruebas miden una combinación de atributos requerida por un criterio complejo, en cuyo caso purificar la prueba puede reducir la cobertura del criterio y disminuir por ende la validez.

La selección de reactivos que aumenten la validez de la prueba relacionada con el criterio puede compararse con la selección de pruebas que produzcan la mayor validez para una batería. Recordará (capítulo 6) que el instrumento que más contribuye a la validez de la batería es el que tiene la mayor correlación con el criterio y la menor correlación con las otras pruebas de la batería. Si se aplica este principio a la selección de reactivos, significa que los más satisfactorios son los que tienen mayor validez externa y menor coeficiente de consistencia interna, de modo que un reactivo que tiene una elevada correlación con el criterio externo pero una correlación relativamente baja con la puntuación total será preferible a uno que correlacione alto tanto con el criterio como con la puntuación obtenida en la prueba, ya que es de suponer que el primero mide un aspecto del criterio que no cubre adecuadamente el resto del instrumento.

Parecería que los reactivos pueden seleccionarse con los mismos métodos empleados para elegir las pruebas que han de incluirse en una batería. Así, cada reactivo debería correlacionarse con el criterio externo y con el resto de los reactivos, y los más adecuados serían ponderados mediante una ecuación de regresión; sin embargo, dicho procedimiento no es factible ni teóricamente aceptable, no sólo por lo agobiante de los cálculos, sino porque, además, las correlaciones entre reactivos también están sujetas a una gran fluctuación del muestreo y los valores resultantes de la regresión serían demasiado inestables para ofrecer una base satisfactoria para la selección de reactivos. Pero una objeción aún más importante es que el contenido de la prueba será demasiado heterogéneo para permitir cualquier interpretación significativa de su puntuación.

Tanto la validez externa como la consistencia interna son objetivos importantes en la elaboración de la prueba, de cuya naturaleza y propósito depende la relativa importancia que se conceda a cada uno. Para muchas finalidades, un compromiso satisfactorio consiste en distribuir los reactivos relativamente homogéneos en pruebas o subpruebas separadas, cada una de las cuales cubre un aspecto distinto del criterio externo. De este modo se obtiene una mayor cobertura al aplicar diversos instrumentos, de los que cada uno produce una puntuación relativamente clara, en lugar de la heterogeneidad de una sola prueba. Con tal procedimiento, los reactivos con un bajo índice de consistencia interna no se eliminan, sino que se separan, lo que permite obtener una elevada consistencia interna dentro de cada subprueba o grupo de reactivos.

Índices estadísticos de la capacidad discriminativa del reactivo. Dado que las respuestas a los reactivos generalmente se registran como correctas o incorrectas, la medición de la discriminación del reactivo suele incluir una variable dicotómica (el reactivo) y una variable continua (el criterio). En ciertas situaciones, también el criterio puede ser dicotómico (como graduarse o no de la universidad o éxito contra fracaso laboral) o bien puede ser dicotomizado si se pretende hacer un análisis.

En la elaboración de pruebas se han desarrollado y empleado más de 50 índices diferentes de la capacidad discriminativa del reactivo. Una diferencia entre ellos tiene que ver con su aplicabilidad a medidas dicotómicas o continuas. Además, algunos suponen que entre los que pueden aplicarse a las variables dicotómicas existe una distribución normal y continua del atributo sobre el que se impuso de manera artificial la dicotomía, mientras que otros presumen que ésta existe verdaderamente. Otra diferencia tiene que ver con la relación entre la dificultad del reactivo y la discriminación. Algunos índices miden la discriminación del reactivo independientemente de su dificultad, en tanto que otros producen un mayor índice de discriminación para los reactivos próximos al grado de dificultad de .50 que para aquellos que se encuentran en los extremos del continuo de dificultad.

A pesar de las diferencias en el procedimiento y las suposiciones, la mayor parte de los índices de la capacidad discriminativa del reactivo arrojan resultados muy similares (Oosterhof, 1976). Aunque el valor numérico de los índices puede diferir, los reactivos que se conservan y los que se rechazan sobre la base de diferentes índices de discriminación son en gran medida los mismos; de hecho, la variación en los datos de la capacidad discriminativa del reactivo de una a otra muestra generalmente es mayor que entre diferentes métodos.

Uso de grupos extremos. Una práctica común en el análisis de reactivos es comparar la proporción de casos que aprueba un reactivo en grupos de criterio contrastantes. Cuando el criterio se mide sobre una escala continua, como sucede con las calificaciones escolares, las calificaciones laborales, los registros de producción o las puntuaciones totales en la prueba, se seleccionan los grupos de criterio superior (S) e inferior (I) de los extremos de la distribución. Aunque la diferenciación será obviamente mayor entre más extremos sean los grupos, el uso de grupos muy extremos (como el 10 por ciento superior e inferior) puede reducir la confiabilidad de los resultados por el pequeño número de casos utilizado. En una distribución normal, el punto óptimo en el que las dos condiciones se equilibran se encuentra en el 27 por ciento superior e inferior (T. L. Kelley, 1939). Cuando la distribución es más plana que la curva normal, el porcentaje óptimo es ligeramente mayor al 27 y se aproxima al 33 por ciento (Cureton, 1957b). Con los grupos pequeños, como los que se encuentran en un salón de clases, el error estadístico del muestreo es tan grande que sólo es posible obtener resultados aproximados; por lo tanto, en estas condiciones no es necesario preocuparse por el porcentaje exacto de casos en los dos grupos de contraste. Cualquier número conveniente entre 25 y 33 por ciento resultará adecuado.

Con las muestras grandes y normalmente distribuidas que se emplean en el desarrollo de las pruebas estandarizadas, se acostumbra trabajar con el 27 por ciento superior e inferior de la distribución del criterio. Muchas de las tablas y ábacos preparados para facilitar el cálculo de los índices de la capacidad discriminativa del reactivo se basan en la suposición de que se ha seguido la "regla del 27". Conforme ha aumentado la dis-

ponibilidad de computadoras de alta velocidad, es probable que muchos de los procedimientos que se desarrollaron para facilitar el análisis de reactivos sean reemplazados gradualmente por métodos más exactos y sofisticados. Con los equipos de cómputo, es mejor analizar los resultados de la muestra entera en lugar de trabajar con los extremos superior e inferior.

Análisis simple con grupos pequeños. Dado que el análisis de reactivos se realiza a menudo con grupos pequeños, como los estudiantes que presentan un examen en el salón de clases, consideremos primero un sencillo procedimiento especialmente adecuado para esta situación. Supongamos que en un grupo de 60 estudiantes hemos elegido a los 20 (33 por ciento) con la mejor puntuación y a los 20 con la más baja. Ahora tenemos tres grupos de exámenes que podemos llamar superior (S), medio (M) e inferior (I). Primero necesitamos contar el número de respuestas correctas dadas a cada reactivo por los estudiantes de los tres grupos. Eso puede hacerse más fácilmente si en una lista se coloca el número del reactivo en una columna y preparamos otras tres columnas con el encabezado S, M e I. Al llegar al examen de cada estudiante, basta anotar una marca al lado de cada reactivo que haya contestado correctamente. El procedimiento se repite con cada uno de los 20 exámenes del grupo S, luego con los 20 exámenes del grupo M y con los 20 del grupo I. Ahora podemos contar las marcas y registrar los totales de cada grupo, como se muestra en la tabla 7.1, en la que se presentan a manera de ejemplo los siete primeros reactivos. Es posible obtener un índice aproximado del valor discriminativo de cada reactivo al restar el número de respuestas correctas al reactivo en el grupo I del número de aciertos en el grupo S. Esas diferencias S – I se presentan en la última columna de la tabla 7.1. Con los mismos datos puede obtenerse una medida del índice de la dificultad del reactivo sumando el número de personas que aprobaron cada reactivo en los tres grupos de criterio (S + M + I).

Tabla 7.1

Procedimiento simple de análisis de reactivos: número de personas que contestan correctamente en cada grupo de criterio

Reactivo	S (20)	M (20)	I (20)	Dificultad (S + M + I)	Discriminación (S – I)
1	15	9	7	31	8
2	20	20	16	56[a]	4
3	19	18	9	46	10
4	10	11	16	37	–6[a]
5	11	13	11	35	0[a]
6	16	14	9	39	7
7	5	0	0	5[a]	5
•					
•					
•					
75					

[a] Reactivos elegidos para revisión.

El análisis de la tabla 7.1 muestra la identificación de cuatro reactivos cuestionables que deben recibir mayor consideración o revisarse en clase. Dos reactivos, el segundo y el séptimo, están señalados porque uno parece ser demasiado sencillo, ya que lo contestaron 56 de 60 estudiantes, y el otro demasiado difícil, pues sólo lo aprobaron cinco. Los reactivos cuarto y quinto, aunque con un grado de dificultad satisfactorio, muestran un valor discriminativo negativo y cero respectivamente. Cuando se comparan grupos de este tamaño también es posible considerar en esta categoría a cualquier reactivo con una pequeña diferencia positiva S – I, de aproximadamente tres o menos. Con grupos mayores podemos esperar que ocurran diferencias más grandes por azar en un reactivo no discriminativo.

El propósito del análisis de reactivos en una prueba construida por un maestro consiste en identificar las deficiencias en la prueba o en la enseñanza. Para diagnosticar el problema a menudo es suficiente revisar con el grupo los reactivos cuestionables. Si la redacción del reactivo fue errónea, se puede revisar o eliminar en exámenes posteriores; sin embargo, el análisis puede demostrar que el reactivo era satisfactorio, pero que el punto o tema examinado no se entendió adecuadamente, en cuyo caso el tema puede repasarse o aclararse. Al reducir la fuente de dificultad a menudo es útil realizar un análisis complementario, como se muestra en la figura 7.2, al menos con algunos de los reactivos elegidos para revisión. Esa tabulación permite encontrar el número de estudiantes en los grupos S e I que eligen cada opción al contestar a reactivos particulares.

Aunque el segundo reactivo se incluyó en la tabla 7.2, es poco más de lo que podemos saber al tabular la frecuencia de cada opción incorrecta, ya que sólo cuatro personas del grupo I y ninguna del grupo S eligieron respuestas erróneas; no obstante, la revisión del reactivo con los estudiantes puede ayudar a determinar si en conjunto éste era demasiado sencillo y por ende de poco valor intrínseco, si algún defecto en su

Tabla 7.2

Análisis de la respuesta a reactivos individuales

Reactivo	Grupo	Opciones de respuesta				
		1	*2*	*3*	*4*	*5*
2	Superior	0	0	0	**20**	0
	Inferior	2	0	1	**16**	1
4	Superior	0	**10**	9	0	1
	Inferior	2	**16**	2	0	0
5	Superior	2	3	3	**11**	2
	Inferior	1	3	3	**11**	2
7	Superior	5	3	5	4	3
	Inferior	0	5	8	3	4
•						
•						
•						

Nota: Las opciones correctas se presentan en negritas.

redacción les reveló la respuesta correcta o si se trata de un buen reactivo sobre un punto o tema bien enseñado y recordado. En el primer caso, probablemente se eliminaría, en el segundo se revisaría y en el tercero se mantendría sin cambio.

Los datos sobre el cuarto reactivo indican que la tercera opción tiene algunas implicaciones no previstas que llevaron a nueve de los mejores estudiantes a preferirla sobre la opción correcta. Ese punto podría aclararse pidiéndoles que expliquen por qué eligieron la opción. En el quinto item, el problema parece estribar en la redacción del cuerpo del reactivo o de la opción correcta, ya que los estudiantes que fallaron se distribuyeron de manera uniforme entre las cuatro opciones equivocadas. El séptimo reactivo es inusualmente difícil, ya que se equivocaron 15 sujetos del grupo S y todos los del grupo I. La ligera agrupación de respuestas en la opción incorrecta tres indica que resulta superficialmente atractiva, en especial para los más susceptibles al engaño. De igual modo, el hecho de que ningún miembro del grupo I haya elegido la opción correcta (la uno) indica que estaba planteada de tal manera que tenía una apariencia errónea, al menos para los poco informados. Por supuesto, estos dos atributos son el ideal de los buenos reactivos para todo examen. La revisión dentro del grupo puede mostrar que el séptimo es un buen reactivo que aborda un tema que pocos individuos aprendieron realmente.

El índice de discriminación. Si se expresan como porcentajes los números de personas que aprueban cada reactivo en los grupos S e I, la diferencia entre ambos ofrece un índice de la capacidad discriminativa del reactivo que puede interpretarse independientemente del tamaño de la muestra particular de la que se obtuvo. Este índice ha sido descrito en repetidas ocasiones en la bibliografía psicométrica (véase, por ejemplo, Ebel, 1979; A. P. Johnson, 1951; Oosterhof, 1976) y ha sido designado como $S - I$, *SII*, *SID* o simplemente D. Se ha demostrado que, pese a su simplicidad, concuerda estrechamente con otras medidas más elaboradas de la capacidad discriminativa del reactivo (Engelhart, 1965; Oosterhof, 1976). El cálculo de D puede ilustrarse si se consulta los datos de la tabla 7.1. Primero se convierte en porcentaje el número de personas que aprueba cada reactivo en los grupos S e I. La diferencia entre los dos porcentajes es el índice de discriminación (D) que se muestra en la tabla 7.3, que puede encontrarse entre +100 y –100. Si todos los miembros del grupo S y ninguno del grupo I aprueban un reactivo, D es igual a 100. Por el contrario, si todos los sujetos del grupo I contestan correctamente el reactivo mientras que ninguno del grupo S lo aprueba, D es igual a –100. Si los porcentajes de quienes aprueban un reactivo en ambos grupos son iguales, D es igual a cero.

Como sucede con los diversos índices de la capacidad discriminativa del reactivo, los valores de D no son independientes de la dificultad del reactivo, pero se inclinan por los grados intermedios de dificultad. La tabla 7.4 muestra el valor máximo posible de D para reactivos con diferentes porcentajes de respuestas correctas. Si 100 o 0 por ciento de la muestra total aprueba un reactivo, no puede haber diferencias en el porcentaje de aprobación en los grupos S e I, por lo que D es igual a cero. En el otro extremo, si 50 por ciento aprueba un reactivo, es posible que lo hayan aprobado todos los casos del grupo S y ninguno de los del grupo I, lo que arrojaría una D igual a 100 ($100 - 0 = 100$). Si lo aprueba el 70 por ciento, el valor máximo de D se ilustra de la siguiente manera: (S) 50/50 =100 por ciento; (I) 20/50 = 40 por ciento; $D = 100 - 40 = 60$. Debe recordarse que, para la mayor parte de los propósitos de examinación, se prefieren los reactivos más próximos al grado de dificultad del 50 por ciento, por lo

Tabla 7.3

Cálculo del índice de discriminación

Reactivo	Porcentaje de aprobación		Índice de discriminación (D)
	Grupo superior	*Grupo inferior*	
1	75	35	40
2	100	80	20
3	95	45	50
4	50	80	−30
5	55	55	0
6	80	45	35
7	25	0	25

Nota: Datos tomados de la tabla 7.1.

que los índices de discriminación que favorecen este grado de dificultad suelen ser apropiados para la selección de reactivos.

Coeficiente phi. Muchos índices de discriminación de reactivos informan de la relación entre el reactivo y el criterio de acuerdo con un coeficiente de correlación; uno de ellos es el coeficiente *phi* (ϕ). ϕ se calcula sobre una tabla cuádruple y se basa en la proporción de casos que aprueban y fallan un reactivo en los grupos criterios S e I. Como todos los coeficientes de correlación, los valores se encuentran entre +1.00 y −1.00. El coeficiente ϕ supone que existe una genuina dicotomía entre la respuesta al reactivo y la variable de criterio. En consecuencia, sólo puede aplicarse a las condiciones dicotómicas en las que se obtuvo y no puede generalizarse a cualquier relación subyacente entre los atributos medidos y el criterio. Como el índice *D*, ϕ está sesgado hacia el nivel medio de dificultad, es decir, produce las mayores correlaciones posibles para las dicotomías cercanas a la fracción 50–50.

El nivel de significancia del coeficiente ϕ puede calcularse fácilmente mediante la relación de ϕ con la chi cuadrada y la razón de la curva normal. Al aplicar esta última

Tabla 7.4

Relación del valor máximo de *D* con la dificultad del reactivo

Porcentaje de aprobación del reactivo	Valor máximo de *D*
100	0
90	20
70	60
50	100
30	60
10	20
0	0

podemos identificar el valor mínimo de ϕ que puede ser estadísticamente significativo en los niveles de .05 o .01 mediante las siguientes fórmulas:

$$\phi_{.05} = \frac{1.96}{\sqrt{N}}$$

$$\phi_{.01} = \frac{2.58}{\sqrt{N}}$$

En estas fórmulas, N representa el número total de casos en ambos grupos de criterio. Así, si hay 50 casos en el grupo S y 50 en el grupo I, N sería igual a 100 y el mínimo ϕ significativo a nivel de .05 sería $1.96 \div \sqrt{100} = .196$. Cualquier reactivo cuyo ϕ alcance o supere .196 estará correlacionado con el criterio con un nivel de significancia de .05.

Correlación biserial. Como último ejemplo de una medida de uso común de la capacidad discriminativa de un reactivo, consideremos la correlación biserial (r_{bis}), que difiere de ϕ en dos aspectos importantes. Primero, r_{bis} supone una distribución continua y normal de los atributos que subyacen tanto a la respuesta del reactivo dicotómico como a la variable de criterio. Segundo, produce una medida de la relación entre reactivo y criterio que es independiente de la dificultad del reactivo.

El cálculo de la correlación biserial emplea la media de las puntuaciones de criterio de quienes aprueban y de quienes fallan el reactivo, la *DE* de todo el grupo de criterio, y la proporción de casos de los que aprueban y los que fallan el reactivo. En casi cualquier libro de estadística se proporcionan las fórmulas para calcular r_{bis} (por ejemplo, Guilford y Fruchter, 1978, pp. 304–306). El error estándar de r_{bis} puede calcularse con una sencilla fórmula que utiliza los términos ya presentes en la fórmula de la propia correlación. Debe agregarse que, cuando se tiene acceso a las computadoras, éstas ofrecen una forma directa de obtener tanto las correlaciones biseriales como sus errores estándares.

TEORÍA DE RESPUESTA AL ÍTEM

Regresión ítem-test. Tanto la dificultad como la capacidad discriminativa de un reactivo pueden representarse simultáneamente en un diagrama de regresión ítem-test. Como ejemplo, consideremos una prueba hipotética de 12 reactivos que requieren respuestas cortas del tipo de respuesta libre, como las pruebas de vocabulario de las escalas de inteligencia de aplicación individual. La tabla 7.5 presenta la proporción de personas en cada nivel de puntuación total que contestó correctamente a cada uno de los dos reactivos. Esos datos se grafican en la figura 7.5.

El grado de dificultad de cada reactivo puede definirse como su umbral del 50 por ciento, como se acostumbra al establecer los umbrales sensoriales en la psicofísica. Esto se ha usado en la figura 7.5 al trazar perpendiculares desde los puntos en los que las curvas de ambos reactivos cruzan la línea del 50 por ciento al eje horizontal, sobre el que se localizan las puntuaciones totales correspondientes. Se demuestra así que las personas con una calificación aproximada de 8 tienen una oportunidad 50–50 de

Tabla 7.5

Datos hipotéticos que ilustran la regresión ítem-test

Puntuación total	Proporción de acierto	
	Reactivo 7	*Reactivo 13*
12	1.00	.95
11	.82	.62
10	.87	.53
9	.70	.16
8	.49	.05
7	.23	.00
6	.10	.00
5	.06	.00
4	.03	.00
3	.00	.00
2	.00	.00
1	.00	.00

aprobar el reactivo siete, igual que las personas con una calificación aproximada de 10 en el caso del reactivo 13. La pendiente de la curva indica la capacidad discriminativa de cada reactivo: entre mayor sea la pendiente mayor será la correlación de la ejecución en el reactivo con la puntuación total y mayor el índice discriminativo. A simple vista, la capacidad discriminativa de los reactivos siete y 13 parece ser aproximadamente similar.

Un análisis de las regresiones reactivo-puntuación, como las ilustradas en la figura 7.5, nos permite ver qué tan eficaz es la función de un reactivo. Esos diagramas no sólo combinan la información respecto a la dificultad y capacidad discriminativa del reactivo, sino que también presentan una imagen de la relación entre la ejecución en el reactivo y la puntuación total; por ejemplo, el reactivo siete muestra una inversión en la medida en que en este caso se encontró una mayor proporción de respuestas correctas entre las personas con puntuación de 10 que entre las personas con calificación de 11. Si los resultados se basaran en una muestra pequeña, la inversión podría ser insignificante, pero ilustra la información que el análisis de reactivo puede traer a la luz.

Sin embargo, cuando se pretende hacer un tratamiento matemático, una evaluación precisa y una selección de reactivos, el diagrama resulta demasiado limitado. La aproximación sirvió como punto de partida para el desarrollo de formas complejas de análisis de reactivos, que empezaron a llamar la atención en los setenta y a inicios de los ochenta. Una razón de su popularidad es sin duda la disponibilidad cada vez mayor de computadoras de alta velocidad, sin las cuales resultarían imposibles las tareas de cálculo requeridas por esos análisis. Con la preparación de programas de cómputo para los modelos propuestos de análisis de reactivos, se hacen posibles las aplicaciones prácticas de esos procedimientos más refinados. Estudiaremos las principales características de esta aproximación en las siguientes secciones.

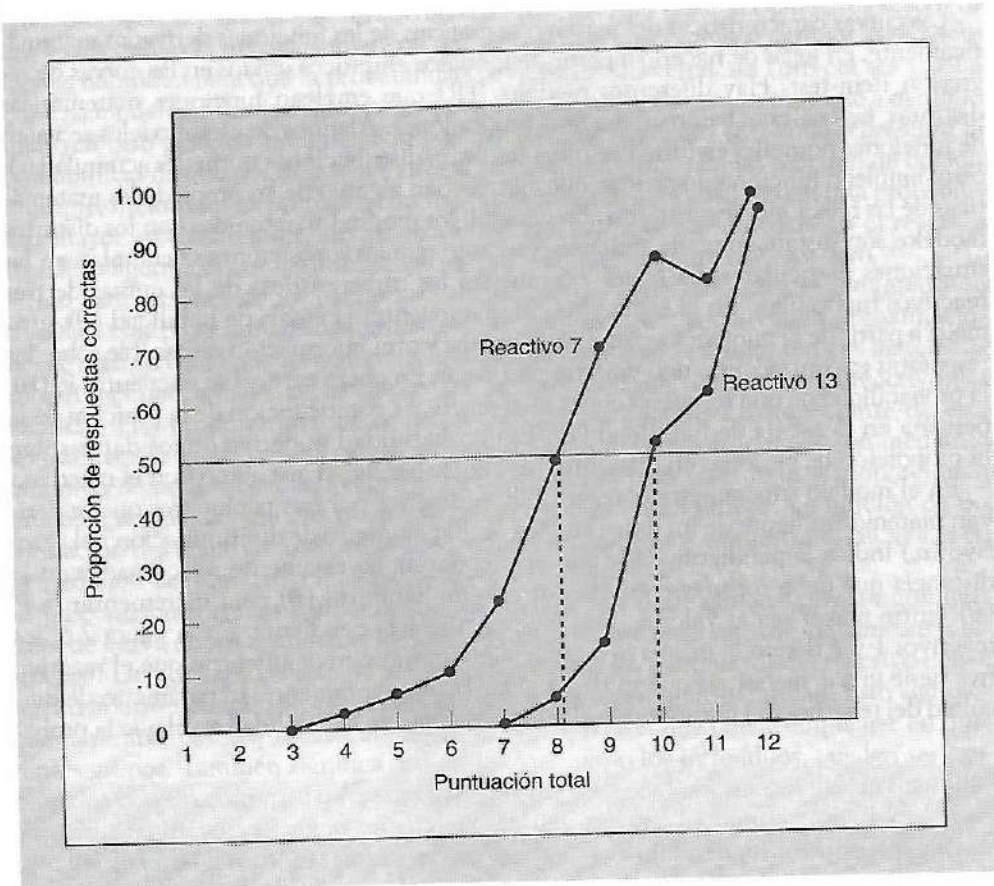


Figura 7.5. Regresión ítem-test para los reactivos siete y 13.
(Datos tomados de la tabla 7.5.)

Teoría de respuesta al ítem (TRI): características básicas.[3] La aproximación matemática considerada como la teoría de respuesta al ítem, también se conoce como teoría de rasgos latentes y teoría de la curva característica del reactivo (CCR). Una peculiaridad de esta aproximación es que relaciona la ejecución en el reactivo con la cantidad estimada del "rasgo latente" en el examinado, simbolizado por la letra griega θ (theta). Tal como se usa en este contexto, un "rasgo latente" se refiere a un constructo estadístico, por lo que no se implica la existencia independiente de una entidad psicológica o fisiológica. En las pruebas cognoscitivas, generalmente se considera que el rasgo latente es la habilidad medida por la prueba. La puntuación total en la prueba a menudo se toma como una estimación de esa habilidad.

[3] Para una revisión clara de la metodología de la TRI y sus aplicaciones, véase Hambleton *et al.* (1991). Puede encontrarse una revisión más técnica y evaluaciones críticas en Hambleton (1989) y en Drasgow y Hulin (1990). Para la introducción de la TRI en la psicometría, véase Lord (1980), D. J. Weiss (1983) y D. J. Weiss y Davison (1981).

Las curvas características del reactivo se grafican de las funciones derivadas matemáticamente, en lugar de hacerlo a partir de los datos empíricos usados en las curvas de regresión ítem-test. Hay diferentes modelos TRI que emplean funciones matemáticas distintas, las cuales se basan en diversas suposiciones. Algunos de esos modelos se valen de funciones normales en forma de ojiva (es decir, distribuciones normales acumuladas); otros emplean funciones logísticas, que aprovechan algunas de las propiedades matemáticas de las relaciones logarítmicas. En general, los resultados obtenidos con los distintos modelos son sustancialmente similares, en tanto que las suposiciones se cumplan en las situaciones particulares. La figura 7.6 muestra las características de las curvas de tres reactivos hipotéticos. En el eje horizontal se encuentra la escala de habilidad (θ), estimada a partir de la puntuación total en la prueba y otra información disponible sobre las respuestas encontradas en una muestra particular. En el eje vertical se encuentra P_i (θ), la probabilidad de una respuesta correcta al reactivo i como función de la posición de la persona en la escala de habilidad (θ). Esa probabilidad se deriva de los datos sobre la proporción de personas en diferentes niveles de habilidad que aprobaron el reactivo.

En el modelo triparamétrico, cada CCR es descrita por tres parámetros que se derivan matemáticamente de los datos empíricos. El parámetro de discriminación del reactivo (a_i) indica la pendiente de la curva, la cual está inversamente relacionada con la distancia que debe recorrerse en el continuo de habilidad (θ) para incrementar la P_i (θ). Entre mayor sea el valor a_i, más inclinada será la pendiente. En la figura 7.6, los reactivos 1 y 2 tienen la misma a_i o capacidad discriminativa, mientras que el reactivo tres tiene una a_i menor, ya que su curva se eleva más lentamente. El parámetro de dificultad del reactivo (b_i) corresponde al punto en el eje de la habilidad en el que la proba-

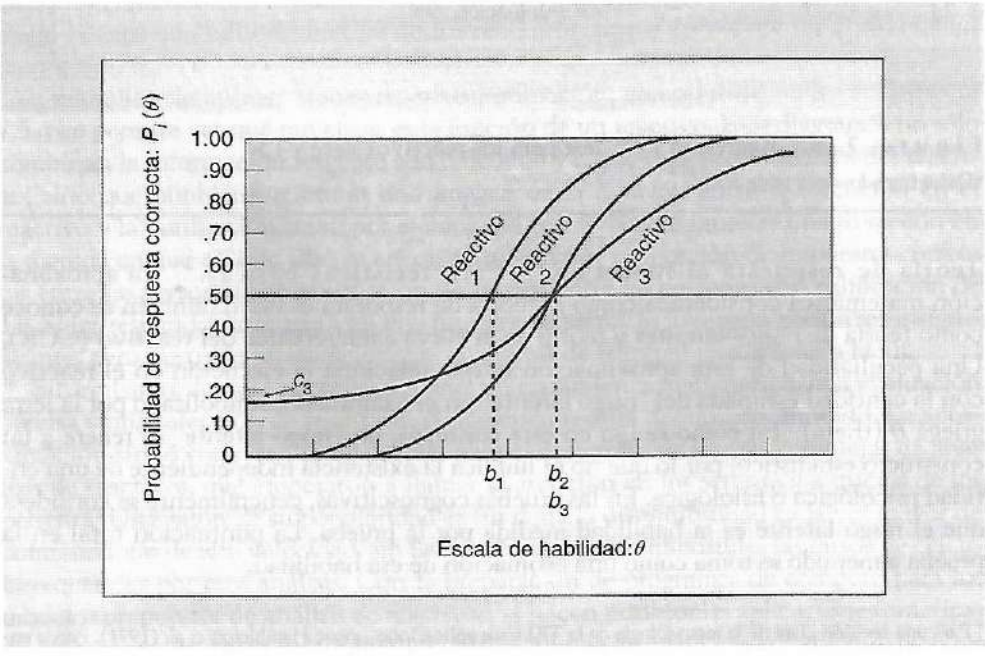


Figura 7.6. Curvas características para tres reactivos hipotéticos.

bilidad de obtener una respuesta correcta, $P_i(\theta)$, es .50. Puede verse que los reactivos dos y tres tienen la misma b_i mientras que el reactivo uno es más sencillo y se requiere menos habilidad para que la probabilidad de obtener una respuesta correcta sea igual a .50. Los modelos TRI diseñados para operar con reactivos de opción múltiple a menudo incluyen otro parámetro, el llamado parámetro de adivinación (c_i),[4] que representa la probabilidad de que por azar se adivine la respuesta correcta. En un reactivo de opción múltiple, incluso los examinados con muy bajo nivel de habilidad tienen una probabilidad mayor a cero de responder de manera correcta, situación ilustrada en el reactivo tres de la figura 7.6, cuya asíntota inferior es considerablemente mayor a cero.

Las estimaciones de los parámetros del reactivo y de la habilidad generalmente se calculan con procedimientos iterativos o de aproximación sucesiva; las aproximaciones se repiten hasta que los valores se estabilizan. Además de arrojar índices matemáticamente perfeccionados de la dificultad y la capacidad discriminativa del reactivo, las técnicas TRI proporcionan otros beneficios. Una característica importante de esta aproximación es su tratamiento de la confiabilidad y del error de medición mediante *funciones de información del reactivo*, que se calculan para cada ítem y proporcionan una base adecuada para elegir los reactivos al elaborar una prueba. La función de información del reactivo toma en cuenta todos los parámetros y muestra la eficiencia en la medición del reactivo a diferentes niveles de habilidad.

La contribución más conocida de los modelos TRI tiene que ver con la naturaleza libre de curva de sus resultados, lo que se describe como *invarianza de los parámetros del reactivo*. Un concepto básico de la teoría TRI es que los parámetros del reactivo deben ser invariantes cuando se calculan en grupos que difieren en habilidad. Esto significa que es posible obtener una escala uniforme de medición que puede emplearse en diferentes grupos. También significa que los grupos, como los individuos, pueden ser examinados con un conjunto diferente de reactivos apropiados a sus niveles de habilidad y que sus puntuaciones serán directamente comparables. La puntuación de cada persona en la prueba no sólo se basa en el número, sino también en los niveles preestablecidos de dificultad de los reactivos contestados correctamente.

Cuando se pretende examinar muchas muestras diferentes, un procedimiento consiste en trabajar con un banco de reactivos que antes haya sido calibrado con una muestra aleatoria grande. Cuando el rango de habilidad es muy amplio, como en las series de pruebas de aprovechamiento que traslapan varios grados escolares, es necesario emplear reactivos comunes (llamados reactivos ancla, de encadenamiento o de calibración) para llenar las brechas entre los grupos. Una vez que se han calibrado los reactivos del banco, puede aplicarse cualquier subconjunto a cualquier grupo o individuo, y las puntuaciones resultantes serán comparables.

Otros modelos TRI. En la sección anterior examinamos un modelo que incluye tres parámetros. Los modelos de dos parámetros, que omiten el parámetro de la respuesta al azar (c_i), son apropiados cuando es posible descartar o minimizar los efectos de la adivinación sobre la ejecución. Rasch (1966, véase también Andersen, 1983) elaboró un modelo de un solo parámetro basado exclusivamente en la dificultad (b_i)

[4] Dado que los modelos de tres parámetros consideran que c_i es independiente de la habilidad, mientras que la adivinación es una función de la misma, algunos investigadores recomiendan que c_i se conozca como una asíntota menor o parámetro de azar para la CCR.

de un conjunto de reactivos, modelo que ha sido adaptado y apoyado por algunos investigadores (por ejemplo, Wright, 1977; Wright y Stone, 1979). El modelo se basa en la suposición de que tanto la adivinación como las diferencias en la capacidad discriminativa de los reactivos son insignificantes. Cuando los exponentes del modelo de Rasch elaboran pruebas, frecuentemente descartan los reactivos que no cumplen con estas suposiciones. Se ha afirmado también que los modelos TRI pueden ser "fuertes" en el sentido estadístico, lo que quiere decir que algunas de sus suposiciones pueden ser transgredidas ilimitadamente sin que se distorsionen los resultados. Esta afirmación, por supuesto, es materia de verificación empírica.

Los modelos considerados hasta ahora suponen que la prueba es *unidimensional*, es decir, que las respuestas pueden atribuirse a un solo atributo. En general, la suposición de unidimensionalidad puede ser adecuadamente satisfecha si la ejecución en la prueba depende de un único atributo predominante, aunque otros rasgos puedan influir en la ejecución de menor manera. Se cuenta también con modelos más generales que pueden aplicarse a pruebas multidimensionales, aunque los procedimientos de cálculo son más laboriosos. Además, se cuenta con otras variantes para manejar respuestas graduadas (más que respuestas dicotómicas o correctas-incorrectas) (Samejima, 1969) o para analizar las diferentes opciones de respuesta en los reactivos de opción múltiple (Bock, 1972).

Estado actual. Aún se debate respecto a los méritos relativos de los modelos TRI, por lo que se necesita una mayor supervisión de los valores derivados matemáticamente, no sólo con datos artificiales y simulación por computadora, sino también con datos reales. En particular se requiere investigar la invarianza de los parámetros del reactivo en situaciones reales; por ejemplo, los mismos reactivos pueden incluir una mezcla diferente de habilidades cuando se presentan a personas con antecedentes diversos o cuando son resueltos por la misma persona en distintas etapas del aprendizaje. Desde otro punto de vista, se dispone ya de una cantidad y una variedad considerable de programas de cómputo para el manejo de los datos TRI (véase, por ejemplo, Hambleton, 1989, pp. 171–172), pero los programas están en un estado constante de cambio debido a las reevaluaciones, revisiones y reemplazos.[5]

Mientras se resuelven las diferencias teóricas y metodológicas, se acude, no obstante, cada vez más a los procedimientos TRI en el desarrollo de pruebas prácticas. Las técnicas de la TRI se incorporan rápidamente a las pruebas recién elaboradas y a las ediciones revisadas de baterías conocidas que ofrecen las editoriales comerciales. Algunos ejemplos son los Tests de Aprovecamiento de California (*California Achievement Tests*) y los Tests Comprensivos de Habilidades Básicas (*Comprehensive Tests of Basic Skills*), así como las Escalas de Habilidad Diferencial (*Differential Ability Scales*) que estudiamos en el capítulo 8. La TRI es particularmente apropiada para ciertos tipos de examinación de reciente aparición, como los Tests Adaptados a las Computadoras, TAC (*Computarized Adaptive Testing*, CAT) que veremos en el capítulo 10. En esa forma de evaluación, el examinado puede responder a un conjunto diferente de reactivos, pero todos se califican sobre una escala uniforme (Wainer *et al.*, 1990). Una aplicación importante

[5] Un ejemplo notable es el programa ASCAL para las calibraciones logísticas TRI para dos y tres parámetros distribuido por Assessment Systems Corporation (cuya dirección aparece en el apéndice B). También es de provecho el reciente desarrollo de la teoría lineal generalizada de respuesta al reactivo (TLGRR), de la que pueden derivarse diferentes modelos TRI que pueden adaptarse a distintos formatos de reactivos (Mellenbergh, 1994).

de la TRI es el proyecto a largo plazo para desarrollar una versión TAC de la Batería de Aptitud Vocacional de las Fuerzas Armadas (Wiskoff y Schratz, 1989).

ANÁLISIS DE REACTIVOS DE LAS PRUEBAS DE VELOCIDAD

Sea la velocidad importante o no para la función medida, los índices calculados de una prueba de velocidad pueden ser engañosos. Salvo por los reactivos que todos o casi todos los examinados tienen el tiempo de intentar, los índices del reactivo que se encuentran en una prueba de velocidad reflejan la *posición* del reactivo en la prueba más que su dificultad o capacidad discriminativa. Los reactivos que aparecen al final de la prueba serán aprobados por un porcentaje relativamente pequeño de la muestra porque sólo unas cuantas personas tendrán el tiempo de llegar a esos reactivos. Independientemente de lo sencillo que puedan ser, e incluso si lo que preguntan son nombres, si se presentan al final de una prueba de velocidad parecerán difíciles y el porcentaje de personas que los resuelva será muy bajo.

De igual modo, los índices de la capacidad discriminativa del reactivo tienden a sobrestimarse en los reactivos que no han sido alcanzados por todos los examinados. Como los individuos más diestros suelen trabajar más rápido, es probable que lleguen a uno de los últimos reactivos en la prueba de velocidad, por lo que, independientemente de la naturaleza del reactivo, si se le presenta al final es posible obtener alguna correlación entre el reactivo y el criterio.

Para evitar algunos de esos problemas, podemos limitar el análisis de cada reactivo a las personas que lo alcanzaron, pero ésta no será una solución del todo satisfactoria a menos que el número de sujetos que no logren llegar al reactivo sea relativamente pequeño. Dicho procedimiento supone el uso de un número de casos que disminuye con rapidez, por lo que los resultados sobre los últimos reactivos serían poco confiables. Más aún, los sujetos en los que se analizan los últimos reactivos probablemente constituyen una muestra seleccionada y por ende no comparable a las muestras mayores usadas para los primeros reactivos. Como ya dijimos, los más rápidos también tienden a ser los más diestros, por lo que los últimos reactivos serían analizados en una muestra superior de individuos. Un efecto de ese factor selectivo podría consistir en la disminución del nivel aparente de dificultad de los últimos reactivos, ya que el porcentaje de los que aprueban sería mayor en el grupo seleccionado superior que en la muestra entera. En ese caso, la aparente dificultad de los reactivos aumenta de manera falsa.

El efecto del procedimiento anterior sobre los índices de la capacidad discriminativa del reactivo es menos evidente, pero no menos real; por ejemplo, se ha observado que los mismos individuos que obtienen bajas puntuaciones tienden a apresurarse cuando presentan la prueba, y marcan los reactivos casi al azar en su esfuerzo por terminar de contestarlos en el tiempo permitido. Como resultado, la muestra en la que se analiza uno de los últimos reactivos probablemente conste de algunos examinados mediocres cuya ejecución no es mejor que el nivel del azar y un número grande de individuos que trabajan de manera rápida y eficiente que tal vez respondan de manera correcta. En este grupo, es probable que la correlación reactivo-criterio sea mayor de lo que resultaría en una muestra más representativa. Por otro lado, al no contar con individuos que contesten al azar, la muestra en la que se analizan los últimos reactivos ha de cubrir un rango de habilidad relativamente estrecho. En esas condiciones, los

índices de discriminación de los últimos reactivos tienden a ser menores que si se calculan sobre una muestra no seleccionada.

Los efectos anticipados de la velocidad sobre la dificultad y la discriminación del reactivo se han verificado empíricamente, tanto al calcular las estadísticas del reactivo con la muestra entera (Wesman, 1949) como cuando se calculan sólo con las personas que intentan resolverlo (Mollenkopf, 1950a). En el último estudio, se aplicaron a grupos comparables de estudiantes de secundaria dos formas de una prueba verbal y dos de una prueba de matemáticas. Cada una de las formas contenía los mismos reactivos que la otra, pero los reactivos que en una aparecían al principio en la otra se colocaban al final. Cada forma se aplicó con un tiempo límite corto (condiciones de velocidad) y con un límite temporal muy libre (condiciones de poder), lo que hizo posible realizar varias comparaciones entre las formas y las condiciones temporales. Los resultados mostraron claramente que la posición de un reactivo en las pruebas de velocidad influye en sus índices de dificultad y discriminación. Cuando el mismo reactivo se presentaba al final de una prueba de velocidad, era aprobado por un porcentaje mayor de quienes intentaban resolverlo, y de esa manera arrojaba una correlación más alta entre reactivo y criterio.

Las dificultades enfrentadas en el análisis de reactivos de las pruebas de velocidad son en esencia similares a las que revisamos en el capítulo 4 al tratar la confiabilidad de estos instrumentos. Para enfrentar esos problemas, se han propuesto varias soluciones, empíricas y estadísticas. Una solución empírica consiste en aplicar la prueba con un límite de tiempo largo para el grupo en que se realizará el análisis de reactivos. Esta solución será satisfactoria siempre que la velocidad no sea un aspecto importante de la habilidad que la prueba pretende medir. Aparte de los problemas técnicos presentados por las pruebas específicas, es conveniente recordar que los datos del análisis de reactivos realizado con las pruebas de velocidad son sospechosos y requieren un escrutinio cuidadoso.

VALIDACIÓN CRUZADA

Significado de la validación cruzada. Es esencial que la validez de la prueba se calcule en una muestra diferente a aquélla en la que se seleccionaron los reactivos. Esta determinación independiente de la validez del instrumento se conoce como validación cruzada. Cualquier coeficiente de validez calculado en la misma muestra que fue empleada para seleccionar los reactivos capitalizará los errores de muestreo aleatorio de esta muestra particular y en consecuencia resultará alto en forma espuria. De hecho, en esas circunstancias puede obtenerse un coeficiente de validez elevado aunque la prueba no tenga validez alguna para predecir el criterio particular.

Supongamos que de una muestra de 100 estudiantes de medicina se elige a los 30 con las mayores calificaciones y a los 30 con las calificaciones más bajas para representar grupos de criterio contrastados. Si ahora se comparan los dos grupos en una serie de rasgos que son irrelevantes para el éxito en la escuela de medicina, indudablemente se encontrarán algunas diferencias al azar. Así, en el grupo de criterio superior puede haber un exceso de personas que provienen de escuelas privadas y de pelirrojos. Si se asignara a cada individuo un punto por comprobar que proviene de una escuela privada y otro punto por ser pelirrojo, las medias de esas calificaciones indudablemente serían mayores en el grupo de criterio superior que en el inferior; sin embargo, esto no es evi-

dencia de la validez de los predictores, ya que el proceso de validación se fundamenta en un argumento circular. En primer lugar, los dos predictores fueron elegidos sobre la base de las variaciones aleatorias que caracterizan a esta muestra particular, y las *mismas* diferencias aleatorias operan para producir las diferencias promedio en la calificación total; sin embargo, cuando se prueban en otra muestra, es probable que desaparezcan o se inviertan las diferencias aleatorias en la frecuencia de graduaciones de escuelas privadas y cabello rojo. En consecuencia, se desplomará la validez de las calificaciones.

Un ejemplo empírico. Una investigación conducida con la prueba de manchas de tinta de Rorschach proporciona una demostración clásica de la necesidad de validación cruzada (Kurtz, 1948). En un intento por determinar la posibilidad de que el Rorschach fuese de utilidad para seleccionar gerentes de ventas para una empresa de seguros de vida, la prueba se aplicó a 80 de esos gerentes que habían sido cuidadosamente elegidos de entre varios cientos de empleados de ocho compañías de seguros, de modo que representaran un grupo de criterio superior de 42 individuos considerados muy satisfactorios por sus compañías y un grupo de criterio inferior de 38 sujetos que se consideraban inadecuados. Los 80 registros de la prueba fueron estudiados por un experto en el Rorschach, que seleccionó un conjunto de 32 signos, o respuestas características, que ocurrían más frecuentemente en un grupo de criterio que en el otro. Los signos que se encontraron más a menudo en el grupo de criterio superior se calificaban como +1 si estaban presentes y 0 si se encontraban ausentes; los que eran más comunes en el grupo inferior se calificaban asimismo como –1 o 0. Como había 16 signos de cada tipo, en teoría las calificaciones totales podían fluctuar entre –16 y +16.

Cuando la clave de calificación basada en esos 32 signos se volvió a aplicar al grupo original de 80 personas, 79 de las 80 se clasificaron correctamente como pertenecientes al grupo superior o inferior. La correlación entre la calificación en la prueba y el criterio sería cercana a 1.00; no obstante, cuando la prueba se validó de manera cruzada en la segunda muestra de 41 gerentes, 21 en el grupo superior y 20 en el inferior, el coeficiente de validez cayó a un insignificante .02, lo que hizo evidente que la clave aplicada en la primera muestra no tenía validez para seleccionar a dicho personal.

Un ejemplo con datos al azar. En un estudio clásico, Cureton (1950) demostró que el uso de una sola muestra para la selección de reactivos y la validación de la prueba puede producir un coeficiente de validez completamente espurio. El criterio a predecir era el promedio de calificaciones de cada uno de 29 estudiantes inscritos en un curso de psicología. El criterio se dividió en calificaciones de B o mayor y calificaciones inferiores a B. Los "reactivos" consistían en 85 etiquetas, que en uno de sus lados tenían un número del uno al 85. Para obtener la calificación en la prueba de cada estudiante se mezclaron las 85 etiquetas en un recipiente y se dejaron caer sobre una mesa. Todas las etiquetas que caían con el lado numerado hacia arriba se registraron como si indicaran la presencia de ese reactivo particular en la ejecución del estudiante. De esa manera se obtuvo un registro completo para cada estudiante en los 29 lanzamientos de las 85 etiquetas, que mostraba la presencia o ausencia de cada reactivo o respuesta signo en esa ejecución del estudiante. Debido al procedimiento seguido para generar esas calificaciones aleatorias, Cureton lo llamó jocosamente *B-Projective Psychokinesis Test* ("Prueba Psicokinética Proyectiva B").

Se realizó un análisis de reactivos con la calificación promedio de cada estudiante como criterio. Sobre esta base, se seleccionaron 24 de los 85 "reactivos", y de éstos, nueve ocurrían con mayor frecuencia entre los estudiantes con una calificación promedio de B o mejor y recibieron un peso de +1; 15 ocurrieron más frecuentemente entre los estudiantes con una calificación menor a B y recibieron un peso de –1. La suma de los pesos de esos reactivos constituyó la calificación total para cada estudiante. A pesar de conocer la derivación aleatoria de esas "calificaciones", su correlación con la calificación de criterio en el grupo original de 29 estudiantes demostró ser de .82. Semejante descubrimiento es similar al obtenido con las puntuaciones del Rorschach en el estudio citado anteriormente. En ambos casos, la aparente correspondencia entre la calificación de la prueba y el criterio fue producto del empleo de las mismas diferencias aleatorias al seleccionar los reactivos y al determinar la validez de las calificaciones totales.

Condiciones que influyen en la disminución de la validez. El monto de la disminución del coeficiente de validez en una validación cruzada depende en parte del tamaño del banco de reactivos original y de la proporción conservada de ellos. Cuando el número original de reactivos es grande y la proporción conservada es pequeña, hay más oportunidad de capitalizar las diferencias al azar y obtener un coeficiente de validez elevado de manera espuria. Otra condición que afecta la magnitud de la disminución en la validación cruzada es el tamaño de la muestra, pues como una validez espuriamente alta en la muestra inicial es resultado de la acumulación de errores del muestreo, los grupos más pequeños (que dan lugar a mayores errores del muestreo) son los que exhiben la mayor disminución de la validez.

La disminución de la validez en la validación cruzada será mínima si los reactivos se eligen a partir de hipótesis derivadas de la teoría psicológica o bien de la experiencia con el criterio. Por ejemplo, si una hipótesis particular requiere que la respuesta "sí" sea más frecuente entre los estudiantes con éxito, entonces el reactivo *no debería* conservarse si los estudiantes *no exitosos* dieran un número significativamente grande de respuestas afirmativas. La aproximación opuesta, del todo empírica, se ejemplifica en el agrupamiento de un conjunto de reactivos sin considerar su importancia para la conducta de criterio para luego conservar los que producen correlaciones positivas o negativas significativas con el criterio. En estas circunstancias, podemos esperar que la disminución sea mucho mayor que en las primeras. T. W. Mitchell y Klimoski (1986) ofrecen una demostración bien diseñada de las diferencias que realmente se presentan en la disminución de la validez cuando los reactivos se eligen mediante estrategias racionales o empíricas. En resumen, la disminución de la validez en la validación cruzada será mayor cuando las muestras son pequeñas, el banco inicial de reactivos es grande, la proporción de reactivos conservados es pequeña y éstos se agrupan sin una lógica previamente establecida.

FUNCIONAMIENTO DIFERENCIAL DE LOS REACTIVOS

Procedimientos estadísticos. El análisis del "sesgo del reactivo" ha recibido mucha atención como parte de la investigación del sesgo de las pruebas para grupos minoritarios. Ese análisis se interesa principalmente en la *dificultad relativa* de ciertos reactivos para grupos que difieren en sus antecedentes culturales o su experiencia. En

la terminología psicométrica, se conoce como funcionamiento diferencial del reactivo (FDR) a esta área de análisis que pretende identificar los reactivos para los cuales personas igualmente capaces de diferentes grupos culturales tienen distintas probabilidades de éxito. Para su identificación se han propuesto muchos métodos, lo mismo estadísticos que procedimientos que incluyen el juicio humano (Berk, 1982; Camilli y Shepard, 1994; Hambleton y Rogers, 1989; P. W. Holland y Thayer, 1988; P. W. Holland y Wainer, 1993; Osterlind, 1983; C. R. Reynolds y Brown, 1984).

Un problema importante son las diferencias en los grupos demográficos (o en otras experiencias) en la dificultad del reactivo que se asocian estrechamente con las diferencias promedio en el nivel de ejecución en el conjunto de la prueba. En consecuencia, es probable que los reactivos que tienen una buena capacidad discriminativa en términos de la calificación total parezcan "sesgados" y por ende se los descarte. Para controlar esas diferencias en la calificación total se han empleado varios procedimientos, y con el acceso creciente a las computadoras, uno de los más prometedores se basa en la teoría de respuesta al ítem (TRI). Esos procedimientos son especialmente apropiados cuando se dispone de muestras grandes. Como ya mostramos, las curvas características del reactivo (CCR) indican para cada uno la probabilidad de obtener una respuesta correcta en relación con la escala de habilidad para la prueba (figura 7.6). Al comparar las CCR de ambos grupos para el mismo reactivo, podemos identificar los reactivos con un considerable funcionamiento diferencial en relación con la ejecución global de los grupos en la prueba, lo que se expresa en una escala uniforme. Los dos reactivos de la figura 7.7 ejemplifican esta comparación; ahí puede verse que para el reactivo uno, las CCR de los grupos A y B son bastante diferentes, mientras que para el reactivo dos son muy similares. En cada caso, puede emplearse el área entre las dos CCR para identificar el rango de habilidad probada en el que hay evidencias de FDR. Una vez identificados los reactivos con funcionamiento diferencial, independientemente del procedimiento empleado, el siguiente paso consiste en averiguar la naturaleza y fuente de la diferencia. La respuesta a esta interrogante determina la disposición del reactivo, para lo cual puede ser necesario contar con procedimientos de juicio combinados con otros análisis estadísticos.

Procedimientos de juicio. No hay un método de análisis de sesgo que sea "el mejor" para todo propósito, y dado que cada uno ofrece una clase diferente de información, es deseable emplear una combinación de ellos. La combinación apropiada depende del uso que quiera hacerse de la prueba y de las inferencias que se pretenda derivar de las puntuaciones. En general, la mejor combinación incluye algunos procedimientos estadísticos y otros que requieren del juicio humano.

Cuando se aplican apropiadamente, los procedimientos de juicio pueden ofrecer información útil que no podría obtenerse de otra manera (Scheuneman, 1982, Tittle, 1982). El análisis mediante el juicio humano es particularmente apropiado en las primeras y últimas etapas de elaboración de la prueba, de modo que preceda y siga al análisis estadístico. Por lo general, se introduce al inicio de la elaboración de la prueba para eliminar el contenido que pueda ser ofensivo o demeritorio de las minorías, o el que perpetúa estereotipos culturales relativos a funciones ocupacionales o sociales. Con este propósito, las principales editoriales procuran que tanto miembros de su equipo como asesores externos representativos realicen regularmente revisiones preliminares de

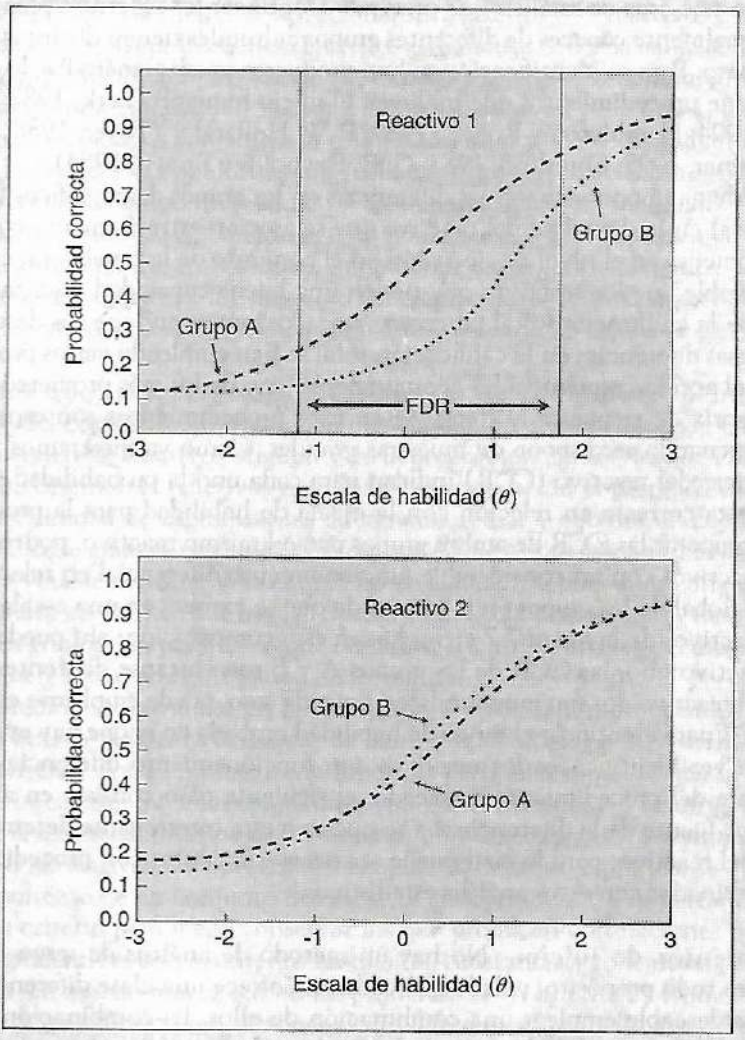


Figura 7.7. Curvas características del reactivo (CCR) que ilustran para dos reactivos un monto grande y uno pequeño de funcionamiento diferencial del reactivo (FDR). (Gráficas adaptadas de Pashley, 1992, con autorización.)

los reactivos (Berk, 1982, capítulo 9). Estas revisiones también ayudan a identificar el contenido de la prueba que puede ser culturalmente restringido y por consecuencia poco familiar para poblaciones particulares de examinados; sin embargo, cabe observar que las revisiones hechas por evaluadores calificados en general *no* han demostrado ser

exitosas para predecir la dificultad relativa o la capacidad discriminativa de los reactivos para diferentes poblaciones (Plake, 1980, Sandoval y Miille, 1980; Scheuneman, 1982). Para cumplir esos propósitos es necesario realizar análisis estadísticos.

Por otro lado, no puede suponerse que todos los reactivos desviados que se identifican mediante procedimientos estadísticos son sesgados. Es necesario que los resultados se interpreten luego de una segunda revisión realizada por expertos, esto permite examinar los reactivos en cuanto a posibles fuentes de variación. Los esfuerzos estadísticos no suelen revelar ninguna característica común ni razón evidente para la desviación, por lo que es necesario considerar cada reactivo de manera individual. Algunos pueden reflejar sencillamente resultados estadísticos producto del procedimiento particular. En otros casos, la ejecución desviada puede resultar de cualquiera de las diversas condiciones que tienen implicaciones distintas para la interpretación de la prueba. La evaluación adecuada de esos reactivos desviados requiere que uno esté familiarizado tanto con el dominio de contenido de la prueba como con las diferencias en la experiencia de las poblaciones examinadas.

Una posible razón de la variación es que un reactivo no mida el mismo constructo en diferentes grupos; por ejemplo, una analogía puede medir razonamiento verbal en un grupo y conocimiento de vocabulario en otro, ya que contiene una palabra clave poco familiar para muchos miembros de un grupo minoritario. De modo similar, un problema aritmético puede medir la habilidad matemática en un grupo y en otro la habilidad de comprender afirmaciones verbales complejas. En ambos ejemplos, el conocimiento que falta no es relevante para el constructo medido por la prueba como un todo. Pero suponga que se descubre que los reactivos matemáticos que incluyen decimales son relativamente más difíciles para los miembros de cierto grupo; esta diferencia sí es relevante para el constructo de la habilidad matemática. Por lo tanto, tales variaciones no son, en este sentido, reactivos sesgados.

Cuando los reactivos sesgados se identifican estadísticamente, la fuente de variación puede aclararse parcialmente por medio de procedimientos estadísticos adicionales, como el análisis de las opciones incorrectas elegidas en un reactivo de opción múltiple. Esos análisis suplementarios en combinación con las revisiones de los jueces, pueden proporcionar la base para la acción adecuada. El reactivo desviado puede descartarse, revisarse o modificarse en su parte defectuosa, las instrucciones de la prueba pueden aumentarse o hacerse más claras, o bien puede decidirse conservar el reactivo original a la luz de las especificaciones de la prueba. El análisis de reactivos puede incluso proponer la reevaluación de las propias especificaciones del instrumento, lo que puede traer como resultado que se modifiquen o se aclaren, dependiendo de las inferencias particulares que puedan sacarse de las calificaciones de la prueba.

Un caso famoso del mal uso del FDR. Se ha dado gran publicidad a lo que probablemente haya sido el primer caso basado en la evaluación del "sesgo de la prueba" mediante el análisis de reactivos que se llevó a una corte legal. El proceso llegó a conocerse como el "caso de la Regla de Oro" por su relación con el uso de un examen de certificación en la contratación de empleados por la compañía de seguros del mismo nombre. La similitud entre el nombre de la compañía y el uso común del mismo en un sentido muy diferente aumentó al parecer la confusión general. El acuerdo final se basó

en la comparación de un grupo de diferencias en un gran porcentaje de personas que aprobaban un reactivo, sin que se hubiera realizado esfuerzo alguno por igualar los grupos en alguna medida de la habilidad que la prueba pretendía evaluar o por considerar la validez de los reactivos para el propósito que el instrumento pretendía cumplir. La decisión se tomó en completa transgresión del concepto de funcionamiento diferencial del reactivo y es probable que eliminara los reactivos que podrían ser los mejores predictores del éxito en el trabajo.

Debido a la magnitud del malentendido y a los precedentes que podría establecer en el uso ocupacional y educativo de las pruebas, la falacia de la decisión de la "Regla de Oro" se examinó desde diferentes perspectivas en evaluaciones críticas (por ejemplo, Lim y Drasgow, 1990), incluida una declaración oficial de la Asociación Estadounidense de Psicología (*American Psychological Association, APA*).[6] La decisión también fue tema de un simposio en la convención anual de la APA, en que la mayor parte de las ponencias se publicaron en un número especial de *Educational and Psychological Measurement: Issues and Practices* (Bond, 1987; Faggen, 1987; Linn y Drasgow, 1987). La consideración de este famoso caso saca a la luz los peligros prácticos de los esfuerzos por evaluar el "sesgo de la prueba" mediante indicadores superficiales e incompletos.

EXPLORACIONES EN EL DESARROLLO DE REACTIVOS

La rápida expansión del uso de computadoras en las décadas de los ochenta y los noventa, en combinación con el progreso en la psicología cognoscitiva, estimularon la investigación de planteamientos innovadores en la elaboración de reactivos. Tradicionalmente, la redacción de reactivos ha sido más un arte que una ciencia, e incluso en las mejores condiciones los redactores de reactivos reciben instrucciones que apenas especifican la forma y contenido que deben cubrir. Sigue siendo práctica común confiar en el examen empírico previo para evaluar el grado de dificultad y la capacidad discriminativa del reactivo. ¿Hay alguna forma de que estas estadísticas puedan inferirse antes del examen mediante el análisis de las propiedades físicas o semánticas de los estímulos? Mejor aún, ¿es posible elaborar reactivos que tengan las características deseadas de grado de dificultad y la capacidad predictiva?, ¿es factible que la manipulación sistemática de las características del estímulo predeterminen las demandas cognoscitivas de los reactivos de la prueba? Se trata de preguntas que estudia la investigación actual mediante procedimientos experimentales y matemáticos (Bejar, 1985, 1991; Carroll, 1987; Embretson, 1985a, 1985b, 1991, 1994, 1995; Freedle, 1990).

Las demandas cognoscitivas de los estímulos de la prueba pueden explorarse mediante el uso de las técnicas de descomposición de tareas desarrolladas por la psicología cognoscitiva. Esos procedimientos permiten investigar las relaciones de distintas características del reactivo con la velocidad y el error de ejecución, y de hecho se han realizado varios estudios con reactivos espaciales (Embretson, 1994; Pellegrino, Mumaw y Shute, 1985). Por ejemplo, los estímulos presentados en las pruebas de analogías espa-

[6] Preparada por el Comité de Pruebas Psicológicas y Evaluación de la APA, la declaración fue aprobada por los comités pertinentes de la APA y por el Consejo de Representantes.

ciales pueden clasificarse de acuerdo con: (1) la complejidad, o número de elementos separados que deben identificarse (por ejemplo, forma, tamaño, posición), y (2) las transformaciones, o número de formas en que se altera al estímulo dentro del par que debe ser evaluado. En algunos problemas de visualización espacial, que requieren que el examinado elija las partes que pueden incluirse para formar un todo determinado, éstas pueden separarse, desplazarse, rotarse o modificarse en alguna combinación.

Otros estudios se han interesado en las características semánticas de los estímulos verbales; por ejemplo, en las pruebas de razonamiento verbal, los reactivos pueden elaborarse de acuerdo con los principios lógicos conocidos (Colberg, 1985; Colberg, Nester y Trattner, 1985; Scheuneman, Geritz y Embretson, 1991; K. Sheehan y Mislevy, 1989; Shye, 1988). Este procedimiento puede asegurar que en el reactivo de muestra sólo una de las opciones de respuesta sea verdaderamente correcta y que diferentes relaciones lógicas estén representadas en una proporción predeterminada. Este procedimiento también puede hacer posible la manipulación de la complejidad lógica del reactivo, cuya relación con el grado de dificultad puede investigarse empíricamente. Algunos investigadores han experimentado con la disposición de series de letras diseñadas para probar el razonamiento inductivo (Butterfield *et al.*, 1985). Para la elaboración sistemática de esa serie de letras primero se desarrolló un conjunto detallado de reglas, luego se formularon hipótesis sobre lo que la gente haría al tratar de entender una serie y se probaron mediante estudios empíricos de la dificultad de series de reactivos de completamiento. .

Embretson (1994) presenta un minucioso análisis y actualización del proceso en el desarrollo de reactivos que comienza con la definición de los constructos que deben evaluarse y continúa con el diseño de un modelo cognoscitivo para la prueba. Los rasgos detallados de este modelo establecen las especificaciones para la redacción de los reactivos, cuya validación empírica permite asegurar qué tan bien se ajustan al modelo en sus aplicaciones prácticas. El procedimiento completo se ilustra en el desarrollo del Test de Habilidad de Aprendizaje Espacial, que no sólo mide la habilidad espacial inicial, sino también su posible modificación después de la instrucción estandarizada.

La investigación sobre la predicción de la dificultad del reactivo a partir de las propiedades físicas y semánticas de los estímulos no sólo facilita la tarea de los redactores en la elaboración de pruebas efectivas, sino que también conduce a la elaboración de reactivos por computadoras. Desde luego, es posible incorporar especificaciones detalladas en programas de cómputo (véase por ejemplo, Butterfield *et al.*, 1985; Embretson, 1994). Sin duda alguna, las ventajas de la evolución de los procedimientos para la preparación de pruebas son enormes, sin embargo es importante no esperar demasiado de cualquier aproximación; por ejemplo, es muy probable que una prueba pueda medir algunos constructos cognoscitivos claramente identificados de manera completa y eficaz y al mismo tiempo carecer de una elevada validez predictiva para ciertos usos prácticos importantes. Por ello es esencial considerar los dos aspectos de validación de constructo que Embretson (1983) designa como representación de constructo y espacio nomotético. La segunda precaución contra la sobregeneralización atañe a la necesidad de conocer el contenido relevante para desempeñarse de manera eficaz en cualquier área temática o campo de experiencia. Los procesos suelen vincularse al contenido, por lo que no pueden evaluarse en forma adecuada en ausencia del contenido apropiado.

En conclusión, cuando los procedimientos innovadores citados en esta sección se aplican apropiadamente, pueden hacer contribuciones importantes a la elaboración sistemática y controlada de los reactivos de las pruebas. Más aún, al identificar los constructos medidos por una prueba, estos procedimientos pueden aumentar de modo considerable nuestra comprensión de por qué algunas pruebas predicen el desempeño en situaciones de criterio. Un beneficio adicional tiene que ver con el uso diagnóstico de la prueba, en la medida en que sea posible vincular a procesos cognoscitivos particulares la fuente de fortalezas y debilidades del individuo. Se trata de metas valiosas, pero su puesta en práctica aún requiere de mucha investigación sobre problemas todavía no resueltos (véase, por ejemplo, Wainer, 1993a). En la actualidad se investiga a fondo la preparación de reactivos que permitan identificar los procesos cognoscitivos que emplean algunos individuos al resolverlos (Willson, 1994). El análisis de los errores que cometen aporta elementos prometedores para este propósito (Kulikowich y Alexander, 1994).

TERCERA PARTE

TESTS de habilidad

Capítulo Ocho

Pruebas individuales

En la segunda parte revisamos los más importantes principios de la evaluación psicológica, ahora estamos listos para aplicarlos a la evaluación de pruebas específicas, es decir, sabemos qué preguntas plantear sobre cada instrumento y dónde buscar las respuestas. Los manuales y el Anuario de Medición Mental (*Mental Measurements Yearbook*, MMY) son algunas de las fuentes principales que pueden consultarse para obtener información sobre cualquiera de los instrumentos citados.[1]

El propósito de lo que resta del libro es doble: por un lado, dar al lector la oportunidad de observar la aplicación de los principios de la examinación en una amplia variedad de pruebas; por otro, familiarizarlo con algunos instrumentos destacados en cada una de las principales áreas. Lo anterior no supone que se pretenda hacer una revisión exhaustiva de las pruebas existentes en cada área, semejante revisión no sólo estaría fuera del alcance de esta obra, sino que antes de su publicación resultaría obsoleta por la rapidez con la que aparecen pruebas nuevas o revisadas. Es por ello que nos concentraremos en la revisión de algunos ejemplos representativos de cada categoría, elegidos por su uso generalizado o porque ilustran desarrollos importantes en el procedimiento de examinación. En la tercera parte consideramos los tests de habilidad, en la cuarta veremos los tests de personalidad y en la quinta las aplicaciones de las pruebas en diversos contextos. A menos que se indique otra cosa, todos los datos sobre los instrumentos que revisaremos provienen del manual o de los complementos técnicos proporcionados por los editores. Los *Estándares de Examinación* (AERA, APA, NCME, 1985) ofrecen directrices más detalladas para ese propósito.

[1] La serie de 10 volúmenes de *Test Critiques* (Keyser y Sweetland, 1984–1994) proporciona otra fuente útil de información y revisiones de centenares de pruebas.

Conocidos tradicionalmente como "tests de inteligencia", los que veremos en este capítulo y el siguiente son descendientes directos de las escalas originales de Binet, éstos se diseñaron para emplearse en diversas situaciones y han sido validados con criterios relativamente amplios (véase L. R. Aiken, 1996). Por lo general ofrecen una sola calificación resumida (como el famoso CI) como índice general del grado de ejecución del examinado, aunque también brindan calificaciones en subpruebas o grupos de subpruebas que evalúan aptitudes más estrechamente definidas. Como muchos de los tests de inteligencia se validan con medidas de aprovechamiento académico, es común que se les conozca como pruebas de aptitud escolar o de inteligencia académica. Generalmente se emplean como instrumentos de clasificación preliminar a cuya aplicación debe seguir la medición de aptitudes especiales, una práctica frecuente al examinar a los adolescentes y adultos normales con fines de consejería educativa, ocupacional, selección de personal y propósitos similares. Otro uso habitual de los tests de inteligencia se encuentra en la evaluación clínica, especialmente en la identificación y clasificación de personas con retardo mental, en cuyo caso es común la aplicación de pruebas individuales. Entre los tests individuales de inteligencia más empleados (en contraste con las pruebas colectivas) están las escalas de Stanford-Binet y de Wechsler, que veremos en este capítulo. Como la de Stanford-Binet es la primera que analizaremos, su revisión es más profunda que la de los otros instrumentos presentados para ilustrar desde el inicio la clase de información que debe considerarse al evaluar un instrumento. Sin embargo, debe quedar claro que los análisis que hacemos en el texto de pruebas particulares no deben considerarse como revisiones, éstas pueden encontrarse en fuentes como el *Anuario de Medición Mental*. Para los propósitos de esta obra, concentramos la atención en las contribuciones especiales de una prueba en particular o en las características que la distinguen de otros instrumentos.[2]

ESCALA DE INTELIGENCIA STANFORD-BINET

Evolución de las escalas. En el capítulo 2 hicimos un breve relato de las escalas originales de Binet-Simon, publicadas en Francia en 1905, 1908 y 1911, de ahí que el lector recuerde que de las muchas traducciones y adaptaciones de los primeros tests de Binet que aparecieron en los Estados Unidos, el más viable fue el de Stanford-Binet.[3] En 1916, se publicó la primera revisión que prepararon Terman y sus colaboradores en la Universidad de Stanford de las escalas Binet-Simon (Terman, 1916). Los cambios y adiciones incluidos en esta revisión fueron tantos que representaron finalmente un nuevo instrumento; por ejemplo, más de una tercera parte de sus reactivos eran nuevos y varios de los anteriores se revisaron, se reasignaron a diferentes niveles de edad o se descartaron; la escala completa se estandarizó con una muestra estadounidense formada aproximadamente por 1 000 niños y 400 adultos; y por último, se proporcionaron instrucciones detalladas para su aplicación y calificación y se empleó el CI por primera vez

[2] Véase el volumen de *Contemporary Intellectual Assessment*, compilado por Flanagan, Genshaft y Harrison (1997), para una excelente revisión de muchos de los temas que estudiamos en la tercera parte de esta obra.

[3] Véase Sattler (1982, 1988) para explicaciones detalladas de las escalas Binet-Simon y del desarrollo, uso e interpretación clínica del Stanford-Binet.

en una prueba psicológica. En su segunda revisión, publicada en 1937, la escala, que constaba de dos formas equivalentes, L y M, fue aumentada y reestandarizada con una nueva muestra de la población estadounidense (Terman y Merrill, 1937). Sin embargo, pese a los esfuerzos por conseguir una muestra representativa de la población, la muestra de 3 184 examinados tenía un nivel socioeconómico superior al de la población de los Estados Unidos, contenía un exceso de casos urbanos y sólo incluía a blancos.

La tercera revisión, publicada en 1960, incorpora en una sola forma (L-M) los mejores reactivos de las dos anteriores (Terman y Merrill, 1960). Al preparar esta versión, los autores enfrentaron un dilema común de los exámenes psicológicos: por un lado, son deseables las revisiones frecuentes porque permiten aprovechar los avances técnicos en la elaboración de las pruebas, beneficiarse de la experiencia en su uso y mantener actualizado el contenido del instrumento. La última consideración es particularmente importante en el caso de los reactivos de información y del material gráfico, ya que pueden verse afectados por los cambios en la moda, aparatos domésticos, automóviles y otros artículos de empleo común, y el uso de pruebas con un contenido obsoleto puede disminuir el *rapport* y alterar el grado de dificultad de los reactivos. Pero, por otro lado, la revisión puede hacer que buena parte de los datos acumulados resulten inaplicables a la nueva forma, es decir, las pruebas que han estado en uso durante muchos años han generado una buena cantidad de material interpretativo cuya utilidad debe sopesarse contra la necesidad de hacer la revisión. Fue por ello que los autores del Stanford-Binet decidieron condensar las dos formas anteriores en una sola, tratando de encontrar un equilibrio entre los riesgos de la obsolescencia y la discontinuidad. La pérdida de una forma paralela no fue un precio demasiado alto por lograr este propósito, ya que en 1960 la necesidad de una forma alterna era menor que en 1937, cuando no se contaba con otra buena escala individual de inteligencia. La escala Stanford-Binet de 1960 *no* incluyó una nueva estandarización de la escala normativa, se emplearon nuevas muestras únicamente para identificar los cambios en la dificultad de los reactivos ocurridos en el periodo intermedio, por lo que en la forma L-M de 1960 las edades mentales y el CI se expresan en términos de la muestra normativa de 1937.

El siguiente paso fue la reestandarización de la forma L-M realizada en 1972 (Terman y Merrill, 1973, Pt. 4). Esta vez el contenido de la prueba prácticamente permaneció inalterado, pero las normas se derivaron de una nueva muestra formada por alrededor de 2 100 casos examinados durante el año académico 1971–1972. En comparación con las normas de 1937, las de 1972 se basaron en una muestra más representativa y, gracias a que fueron actualizadas, pudieron reflejar cualquier efecto de los cambios culturales ocurridos en el ínter sobre la ejecución en la prueba. Es interesante advertir que las últimas normas mostraron cierta mejora en la ejecución, que si bien se observó en todas las edades fue sustancial en la preescolar, que promedió 10 puntos de CI. Entre otros cambios culturales, los autores atribuyeron esta mejora al impacto de los medios de comunicación en los niños pequeños y a la creciente alfabetización y nivel educativo de los padres. Se observó una mejora pequeña pero clara a partir de los 15 años que, como los autores proponen, puede estar asociada a que la proporción de estudiantes que concluyó la secundaria fue mayor en los setenta que en los treinta. Sobre la base de comparaciones longitudinales y transversales, R. L Thorndike (1977) examinó una vez más los cambios normativos y propuso la contribución de otros factores, como la introducción de programas de televisión diseñados para estimular el desarrollo intelectual de los niños de edad preescolar.

El aumento que tuvo lugar en las normas de la década de los treinta o los cuarenta a la de los setenta también se encuentra en otras pruebas de nivel intelectual general (Flynn, 1984, 1987). Desde el punto de vista del usuario, una implicación importante de estos descubrimientos es que parece haber una disminución en la habilidad de los individuos o grupos examinados con las primeras y las últimas formas, porque en las últimas se evalúa su ejecución en comparación con un criterio más elevado. El examinador debe estar al tanto de este posible artefacto al interpretar las calificaciones.

Descripción general de la cuarta edición del Stanford-Binet (SB-IV) La edición actual de esta reconocida escala representa su revisión más extensa (Delaney y Hopkins, 1987; Thorndike, Hagen y Sattler, 1986a, 1986b), que a la vez que conserva las ventajas principales de las ediciones anteriores como un instrumento clínico de aplicación individual, refleja los avances tanto en las conceptuaciones teóricas de las funciones intelectuales como en la metodología para la elaboración de pruebas. La continuidad con las ediciones anteriores se logró en parte al conservar muchos de los tipos de reactivos de las formas previas, pero más importante aún resultó conservar los procedimientos de examinación adaptativa, en la que cada individuo presenta sólo aquellos reactivos cuya dificultad es apropiada para el nivel de ejecución que demuestra.

Al mismo tiempo, el contenido cubierto se extendió más allá de la orientación de lo predominantemente verbal de las formas anteriores para cubrir de manera adecuada las tareas cuantitativas, espaciales y de memoria a corto plazo. El instrumento actual se diseñó para aplicarse de los dos años a la edad adulta, y dado que cada tipo de reactivo se utiliza en un rango de edad tan amplio como es posible, se facilita la comparación en la evaluación entre niveles de edad.

Aplicación y calificación. En la figura 8.1 se muestran los materiales estándares necesarios para aplicar el Stanford-Binet: cuatro folletos de tarjetas impresas que permiten presentar los reactivos al darles vuelta, objetos como cubos, un tablero de formas, cuentas de formas y colores diferentes, así como una lámina grande de una muñeca unisex y multiétnica, una forma de registro, una calificación de las respuestas y una guía para aplicar y calificar la escala.

Como sucede con la mayor parte de los tests de inteligencia individuales, el Stanford-Binet requiere un examinador muy capacitado. Su aplicación, calificación y la interpretación de sus resultados exigen entrenamiento especial y experiencia, pues la familiaridad y la práctica con la escala favorecen la fluidez en la ejecución mientras que las dudas y las chapuzas pueden ser desastrosas para el *rapport*, especialmente con los más jóvenes. Una complicación adicional surge de que los reactivos tienen que calificarse conforme se aplican, ya que el curso subsecuente del examen depende de la ejecución del individuo en los reactivos anteriores.

Con el paso del tiempo, los clínicos llegaron a considerar el Stanford-Binet y a otras escalas similares no sólo como pruebas estandarizadas, sino también como entrevistas clínicas. Las mismas características que dificultan su aplicación también ofrecen la oportunidad de interacción entre examinador y examinado y brindan otras señales para el clínico experimentado. La escala Stanford-Binet y otras técnicas descritas en este capítulo hacen posible observar los métodos de trabajo del individuo, su forma de aproximarse a la solución de problemas y otros aspectos cualitativos de la ejecución. Por su parte, el examinador puede tener la oportunidad de juzgar ciertas caracterísi-

Figura 8.1. Materiales empleados en la aplicación de la escala de Inteligencia Stanford-Binet, cuarta edición.

(Copyright © 1986 de la Riverside Publishing Company. Reproducido con permiso del editor).

cas emocionales y motivacionales, como la concentración, el nivel de actividad, la confianza en sí mismo y la persistencia. Por supuesto, cualquier observación cualitativa hecha en el curso de la aplicación de escalas individuales debe ser reconocida como tal y no interpretarse de la misma manera que las calificaciones de las pruebas objetivas. El mérito de esas observaciones cualitativas depende en gran medida de la capacidad, experiencia y sagacidad psicológica del examinador, así como de su conciencia de las fallas y limitaciones inherentes a este tipo de observaciones.

A diferencia del agrupamiento de edad de las ediciones anteriores, en el SB-IV los reactivos de cada categoría se asignaron a pruebas separadas en orden creciente de dificultad. Se eligieron 15 instrumentos para representar las cuatro áreas cognoscitivas principales: razonamiento verbal, razonamiento abstracto/visual, razonamiento cuantitativo y memoria a corto plazo (véase la figura 8.2). Aunque se agrupan bajo esos cuatro encabezados para propósitos de calificación, las 15 técnicas se aplican en una secuencia mixta para mantener el interés y la atención del examinado. Como puede observarse en la figura 8.2, el grado de dificultad de seis de las pruebas abarcan todo el rango de edad de la escala, mientras que, por la naturaleza de las tareas, las nueve restantes empiezan en un nivel superior o terminan en un nivel menor.

El proceso de aplicación de la cuarta edición del Stanford-Binet consta de dos etapas. En la primera, el examinador presenta la Prueba de Vocabulario que funge como

instrumento de ruta para seleccionar el *nivel de inicio* de las restantes. El punto de inicio de la Prueba de Vocabulario depende únicamente de la edad cronológica del individuo, mientras que en las otras el nivel de inicio se encuentra en una carta que combina la calificación obtenida en Vocabulario y la edad cronológica. En la segunda etapa el examinador debe establecer para cada prueba un *nivel basal* y un *nivel de techo* de acuerdo con la ejecución del individuo. El nivel basal se alcanza cuando se aprueban cuatro reactivos en dos niveles consecutivos, y si esto no ocurre en el nivel de entrada, la prueba continúa hacia abajo hasta alcanzarlo. El nivel del techo se alcanza cuando se fallan tres de cuatro reactivos, o los cuatro, en dos niveles consecutivos. En ese punto se descontinúa la aplicación de esta prueba particular para el individuo.

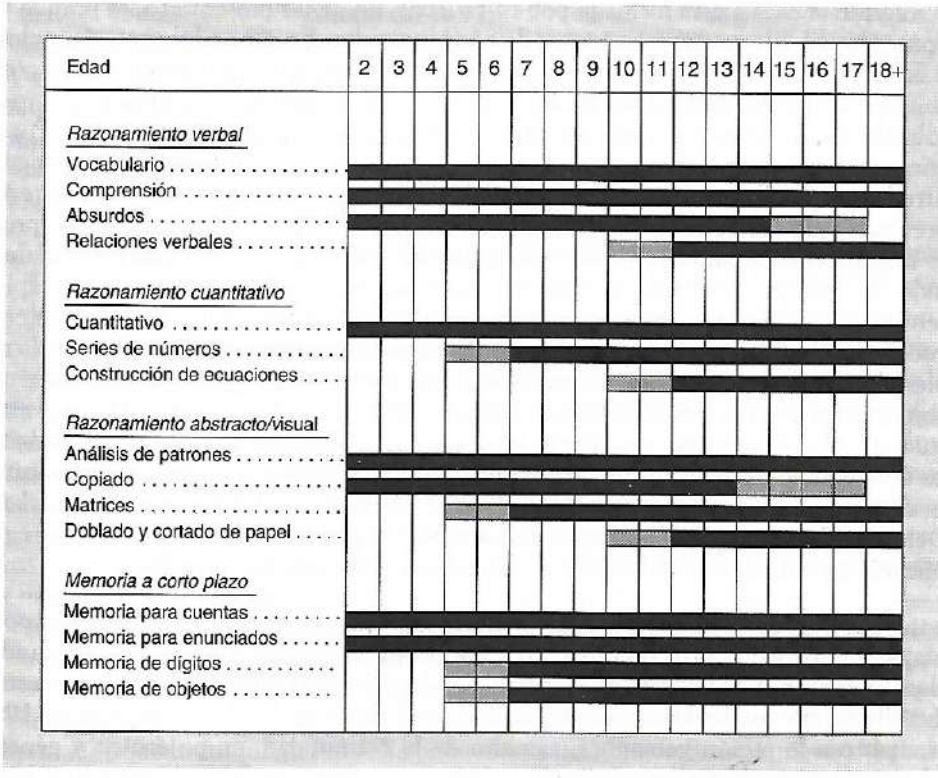


Figura 8.2. Intervalo de edad de las 15 pruebas del Stanford-Binet: Cuarta Edición. *Nota sobre las áreas grises:* En las nueve pruebas diseñadas para rangos restringidos de edad, algunos miembros de la muestra de estandarización que caían fuera de esas edades presentaron la prueba de ruta debido a una ejecución inusualmente alta o baja durante la prueba. Aunque sus calificaciones se emplearon para estimar la ejecución de la muestra de edad para las tablas normativas, estas estimaciones se incluyeron con una advertencia especial acerca de su uso. Para detalles, véase *Guide* (Thorndike *et al.*, 1986a, p. 7) y *Technical Manual* (Thorndike *et al.*, 1986b, p. 30).

(Adaptado de *The Stanford-Binet Intelligence Scale: Fourth Edition, Guide for administering and scoring*, p. 7. Copyright © 1986 de la Riverside Publishing Company, Reproducido con autorización del editor.)

A medida que aplica cada reactivo, el examinador registra la calificación en la forma correspondiente. La puntuación cruda de cada prueba se encuentra al registrar el número del reactivo más elevado que se haya aplicado y restarlo del número total de reactivos que se intentaron y fallaron. Once de las pruebas tienen también reactivos de muestra que no se incluyen en la calificación porque su único propósito es facilitar la familiarización. En la mayor parte de las pruebas, cada reactivo tiene una sola respuesta correcta que el examinador puede consultar en los libros de reactivos y en la forma de registro. Todos los reactivos son aprobados o fallados según las normas especificadas. Cinco de las pruebas requieren respuestas libres y por ende exigen directrices más amplias de calificación que se proporcionan en la *Guía*.[4] En la guía de calificación también se cubren algunas respuestas ambiguas que requieren de mayor indagación por parte del investigador.

Aunque la escala está formada por 15 pruebas, nadie las presenta todas porque algunas sólo son adecuadas para rangos de edad limitados. En general, la batería incluye de ocho a 13 pruebas, según la edad del examinado y su ejecución sobre la ruta o camino de la prueba. Se supone que la aplicación dura entre 30 y 90 minutos, pero es probable que los examinadores con menor experiencia requieran más tiempo. Por lo general es posible aplicar la escala en una sola sesión con unos minutos de intermedio entre las pruebas. Para ciertos propósitos, la *Guía* propone el uso de algunas baterías abreviadas que requieren menos tiempo de examinación y se concentran en las pruebas más adecuadas para el objetivo particular del examen. Incluyen una batería abreviada de seis pruebas para propósitos generales y una de cuatro para la rápida identificación, ambas tienen al menos una técnica de cada una de las cuatro áreas cognoscitivas. También se sugieren tres baterías (apropiadas para cada uno de los tres niveles de edad) para evaluar a estudiantes que pretenden ingresar a programas de sobredotados y otras tres (también correspondientes a los tres niveles de edad) para estudiantes que presentan problemas de aprendizaje en la escuela. En todas las baterías abreviadas se siguen los procedimientos estandarizados para determinar los niveles de inicio asi como para la aplicación y calificación. El *Examiner's Handbook* (Delaney y Hopkins, 1987) aclara muchos temas de procedimiento concernientes a la aplicación y calificación del SB-IV con diferentes clases de examinados.

Estandarización y normas. La muestra de estandarización constó de un poco más de 5 000 casos entre los dos y los 23 años, examinados en 47 estados (incluidos Alaska y Hawai) y el Distrito de Columbia en los Estados Unidos. La muestra se estratificó de forma que igualara las proporciones del censo realizado en ese país en 1980 en cuanto a la región geográfica, tamaño de la comunidad, grupo étnico y género. Además, se supervisó el nivel socioeconómico en términos de la ocupación y nivel educativo de los padres. Los resultados revelaron cierta sobrerrepresentación de los niveles superiores y subrepresentación de los inferiores. Esas discrepancias se ajustaron mediante la ponderación diferencial de las frecuencias al calcular los valores de las puntuaciones en las tablas normativas. De modo que cada individuo de antecedentes socioeconómicos altos contaba como parte de un caso, mientras que el individuo de posición socioeconómica baja contaba como más de uno.

Las tablas normativas se utilizan para convertir las puntuaciones no procesadas o directas de cada una de las 15 pruebas en Calificaciones Estándares por Edad (CEE).[5]

[4] Las cinco pruebas en cuestión son Vocabulario, Comprensión, Absurdos, Copiado y Relaciones verbales.

Se trata de calificaciones normalizadas dentro de cada grupo de edad con una media de 50 y una *DE* igual a ocho. Para los dos y los cinco años se proporcionan tablas normativas con intervalos de cuatro meses, a intervalos de seis meses entre los seis y los 10 años, e intervalos de un año entre los 11 y los 17 años, y una única tabla normativa para el nivel de los 18 a los 23. Para cada una de las pruebas aplicadas, la forma de registro proporciona una carta para trazar el perfil de la ejecución CEE del examinado.

También se cuenta con Calificaciones Estándares por Edad para cada una de las cuatro áreas cognoscitivas y para la ejecución compuesta en toda la escala. Estas CEE por área y compuesta pueden encontrarse en las tablas normativas de las pruebas aplicadas a cada individuo. Las CEE también son calificaciones estándares normalizadas, con una media de 100 y una *DE* de 16, por lo que (aunque el uso del término CI se haya abandonado por completo) se expresan en las mismas unidades que los CI de desviación de las primeras ediciones del Stanford-Binet. Para propósitos especiales, existen condiciones que permiten que el examinador encuentre las CEE de cualquier combinación deseada de calificaciones de dos o más áreas ("compuestos parciales"). Por ejemplo, la combinación de razonamiento verbal y cuantitativo corresponde bien a la "aptitud escolar" y puede ser de gran interés para la evaluación del aprovechamiento o habilidad académica.

Confiabilidad. Dado que no se cuenta con una forma alterna de la cuarta edición del Stanford-Binet, sólo puede evaluarse la confiabilidad mediante los procedimientos de consistencia interna o test-retest. La mayor parte de los análisis emplearon la técnica de Kuder-Richardson, misma que se aplicó en los datos de toda la muestra de estandarización. Como era de esperarse, la puntuación compuesta obtenida en la batería produjo elevados coeficientes de confiabilidad en todos los niveles de edad, fluctuando de .95 a .99. La confiabilidad para cada área cognoscitiva también fue alta, y aunque variaba según el número incluido de pruebas para cada área, iba de .80 a .97. En el caso de las pruebas separadas, la confiabilidad estaba arriba de .80 y en la parte inferior de .90, salvo en el caso de Memoria de Objetos, una prueba corta de 14 reactivos cuya confiabilidad fluctúa entre .66 y .78. En general, la confiabilidad tendía a ser más alta en los niveles superiores de edad que entre los más jóvenes.

Se obtuvieron datos suplementarios sobre la confiabilidad test-retest en 57 preescolares de cinco años y 55 escolares de ocho años a los que se volvió a aplicar la prueba después de intervalos de dos a ocho meses. En general, la estabilidad era buena para la puntuación compuesta, con coeficientes de confiabilidad de .91 y .90 para los dos grupos. Aunque la confiabilidad de la puntuación del área de Razonamiento Verbal es superior a .80, las puntuaciones de otras áreas y de las pruebas separadas revelaron una considerable fluctuación. Es difícil interpretar esos descubrimientos debido a la posible influencia de los rangos restringidos en algunas pruebas y a un perceptible efecto de la práctica que puede variar sustancialmente entre individuos.

Además de los coeficientes de confiabilidad, tanto en la *guía* como en el *manual técnico* puede encontrarse el error estándar de medición (EEM) de cada nivel de edad para cada instrumento, de la puntuación de área y de la compuesta. Estos EEM ayudan a

[5] Las tablas se encuentran en Thorndike *et al.*, 1986a, pp. 183–188. Se estimaron unos cuantos valores CEE, basados en menos de 100 casos, para todo el conjunto de edad y se sombrean en las tablas. Estas calificaciones se obtuvieron cuando los individuos tenían una ejecución desusualmente elevada o baja en la prueba de ruta relativa a su edad (Thorndike *et al.*, 1986b, pp. 29–30).

evaluar las puntuaciones individuales y a interpretar las diferencias entre calificaciones para realizar el análisis del perfil. La CEE compuesta global (M = 100, *DE* = 16) tiene un EEM de alrededor de dos a tres puntos; por ejemplo, si tomamos 2.5 como un valor intermedio aproximado, esto significa que hay una posibilidad de dos a uno de que la puntuación "verdadera" compuesta de un individuo difiera por 2.5 puntos o menos de la puntuación obtenida y hay 95 de 100 posibilidades de que no varíe por más de cinco puntos ($2.5 \times 1.96 = 4.90$).

El *manual del examinador* del Stanford-Binet (Delaney y Hopkins, 1987) ofrece un marco interpretativo que busca generar y realizar la validación cruzada de las hipótesis derivadas de los datos cuantitativos y cualitativos que se obtuvieron con la batería. El análisis cuantitativo sigue un modelo sugerido inicialmente por F. B. Davis (1959) y aplicado, entre otros, por Kaufman (1979, 1994) a las escalas de Wechsler, que en esencia consta de comparaciones entre la puntuación compuesta y las puntuaciones de las cuatro áreas (véase la figura 8.2), en las que se evalúa la significancia estadística de las diferencias en términos del EEM. La frecuencia de las diferencias obtenidas se compara con los datos de la tasa base de la muestra de estandarización. Además, es posible evaluar sistemáticamente las fortalezas y debilidades intraindividuales en las habilidades que cubre cada instrumento al comparar la ejecución promedio de una persona en las puntuaciones compuestas y de las áreas con las puntuaciones en las pruebas individuales. El manual proporciona la información necesaria para realizar este análisis del perfil y cuatro ejemplos de su aplicación, lo que ha demostrado ser de gran valor lo mismo para los examinadores principiantes que para los más experimentados.

Validez. De acuerdo con el concepto actual de validación, en el desarrollo de la cuarta edición de la escala de Stanford-Binet se siguieron diversas aproximaciones para identificar y definir los constructos que deberían incluirse. La elección inicial de los constructos se realizó en función de la información disponible sobre la naturaleza y medición de la inteligencia (R. L. Thorndike *et al.*, 1986b, capítulo 1). La experiencia con las primeras ediciones de la propia escala de Stanford-Binet, con la fortaleza y las debilidades reveladas, orientó los planes y decisiones en la elaboración de la prueba; por ejemplo, la separación de los tipos de reactivos en subpruebas confiables constituyó un sustituto necesario de la práctica clínica tradicional de realizar un análisis informal de patrones mediante el agrupamiento subjetivo de reactivos.

Después de elegir y definir de modo preliminar los constructos que se pretendía evaluar con el SB-IV, se identificaron los reactivos afines a la definición de los constructos y se desarrollaron otros que también lo fueran. Para realizar el análisis de reactivos se sometió al banco de ítemes a procedimientos estadísticos complejos, incluyendo la evaluación del sesgo por medios estadísticos y el empleo de jueces (R. L. Thorndike *et al.*, 1986b, capítulo 2). La versión final de la escala resultante de los intentos preliminares y de algunos ensayos de campo se aplicó a la muestra de estandarización y además se investigaron los tres tipos principales de datos de validación: (1) la intercorrelación y el análisis factorial de las puntuaciones, (2) la correlación con otros tests de inteligencia, y (3) la comparación de la ejecución de grupos excepcionales previamente identificados (Thorndike *et al.*, 1986b, capítulo seis).

Primero, en cada nivel de edad de la muestra de estandarización, se encontraron las intercorrelaciones entre las puntuaciones de todas las pruebas, áreas y puntuaciones compuestas, y se emplearon las correlaciones medianas de todas las edades en un aná-

lisis factorial confirmatorio. El propósito principal de dicho análisis era someter a prueba la hipótesis de que existe un factor general que da cuenta de las correlaciones entre las pruebas de distintas áreas y factores de grupo que explican las correlaciones residuales para cada área. En cada uno de los tres grupos de edad (de dos a seis, de siete a 11 y de 12 a 18–23) se condujeron análisis factoriales similares con las correlaciones intermedias.

Los resultados de los tres análisis factoriales revelaron cargas sustanciales de un factor general en todas las pruebas, lo que justificaría el uso de una calificación compuesta global. En tres de las cuatro áreas, los factores de grupo explicaban una proporción considerable de la varianza residual común del área. La excepción fue el área de Razonamiento Abstracto/Visual, en la que se encontró un elevado grado de especificidad entre las cuatro pruebas. Podría especularse que la imposibilidad de encontrar evidencias claramente definidas de la existencia de un factor de grupo en esta área puede estar asociada con los efectos acumulativos de un programa de estudios que no está tan organizado en relación con el contenido espacial-perceptual como lo está con respecto a contenidos verbales y numéricos. Las experiencias personales cotidianas que contribuyen al desarrollo de habilidades espaciales y perceptuales no se encuentran organizadas sistemáticamente en "cursos" o áreas de contenido como las experiencias académicas, por lo que es menos probable que se fomenten patrones comunes de relaciones entre individuos (Anastasi, 1970, 1986b).

El uso de una calificación compuesta como medida de habilidad intelectual general está apoyada por los resultados de los análisis factoriales reportados en los manuales de los análisis conducidos independientemente por otros investigadores con los datos de estandarización del SB-IV (R. M. Thorndike, 1990); sin embargo, los investigadores no están de acuerdo en el número y naturaleza de factores más reducidos (véase también McCallum, 1990). La situación se complica porque el SB-IV consta de diferentes conjuntos de tests para distintas edades, por lo que los datos crudos difieren para los análisis factoriales (es decir, las correlaciones entre las puntuaciones de la prueba). De lo anterior resulta que en los diferentes niveles de edad aparecen distintos tipos y número de factores, que pueden fluctuar entre dos y cuatro. La confusión aumenta por el hecho de que en las diferentes investigaciones se aplica de diversas formas el análisis factorial; sin embargo, conforme aumenta la edad el modelo de cuatro factores propuesto en el desarrollo del SB-IV se aproxima más cercanamente, en especial cuando en lugar de análisis factoriales exploratorios se realizan análisis factoriales confirmatorios.

La segunda fuente de datos de validación se basó en un conjunto de estudios realizados con grupos a los que se aplicaron la cuarta edición del Stanford-Binet y otro test de inteligencia, incluyendo la forma L-M del propio Stanford-Binet.[6] Se incluyeron cuatro grupos, uno formado por niños inscritos en grupos regulares y designados como "no excepcionales" por sus escuelas y tres grupos "excepcionales" cuyos integrantes participaban respectivamente en programas especiales para estudiantes sobredotados, así como para estudiantes con problemas de aprendizaje y para individuos con retardo mental. Como era de esperar a partir de las semejanzas y diferencias en el contenido de ambas ediciones, en la muestra no excepcional se encontró una correlación de .81 entre el CI de desviación de la versión anterior del Stanford-Binet (Forma L-M) y la calificación compuesta en la cuarta edición; la siguiente correlación más alta fue con el

[6] Los otros instrumentos fueron el WISC-R, el WAIS-R, el WPPSI, y la K-ABC, que estudiamos en este capítulo.

Área de Razonamiento Verbal (.76), mientras que la más baja fue con el Área de Razonamiento Abstracto/Visual (.56). En todos los grupos las correlaciones entre la calificación compuesta y las calificaciones por área del SB-IV con las calificaciones totales o parciales obtenidas en otros tests de inteligencia generalmente fueron coherentes con las hipótesis relativas a los constructos evaluados. Al mismo tiempo, un examen detallado de todas las correlaciones entre calificaciones particulares en la cuarta edición y calificaciones particulares en otros tests de inteligencia contribuye a una mejor comprensión de los constructos medidos ahora por el Stanford-Binet.

El tercer conjunto de estudios especiales realizados en muestras excepcionales demostró que el SB-IV permite identificar apropiadamente el nivel de ejecución de individuos sobredotados, con problemas de aprendizaje y con retardo mental. La muestra de sobredotados obtuvo promedios significativamente superiores a la muestra de estandarización en la calificación compuesta y en las calificaciones de cada una de las cuatro áreas. Las medias de los estudiantes con problemas de aprendizaje y mentalmente retardados en las calificaciones compuesta y por áreas fueron significativamente menores a las medias de la muestra de estandarización, y las medias de los examinados con retardo fueron significativamente inferiores a las de los sujetos con problemas de aprendizaje. Debe añadirse que en todos los estudios realizados con grupos excepcionales, los participantes habían sido previamente identificados sobre la base del uso de pruebas u otros índices de ejecución y que en ningún caso se incluyó el SB-IV en esta identificación.

Una revisión subsecuente de la investigación de la validez de la escala SB-IV (Laurent, Swerdlik y Ryburn, 1992) llegó a la conclusión de que parece ser una medida al menos tan buena de la habilidad intelectual general como otras pruebas disponibles, que tiene una buena correlación con las medidas de aprovechamiento y que puede diferenciar entre las personas con retardo, las sobredotadas y las que tienen daño neurológico. Los revisores proponen que el SB-IV puede ser el instrumento de selección a elegir para la evaluación de niños sobredotados porque en todas las pruebas proporciona un techo elevado para todos los niveles de edad; pero también señalan que la prueba carece de reactivos de piso que sean lo suficientemente bajos como para poder diagnosticar el retardo mental entre los más jóvenes.

Se ha acumulado con rapidez la investigación necesaria para fortalecer el significado interpretativo de las calificaciones de varias pruebas y de las combinaciones de calificaciones del SB-IV. Además, en Sattler (1988), Glutting y Kaplan (1990) y Kamphaus (1993) puede encontrarse una orientación comprensiva para su uso. La actual revisión del Stanford-Binet constituye un avance genuino en la elaboración de la escala, que además de ofrecer a los examinadores la flexibilidad para evaluar habilidades separadas que son apropiadas para finalidades específicas, concuerda estrechamente con los descubrimientos tanto teóricos como empíricos sobre la naturaleza de la inteligencia (véase el capítulo 11).

LAS ESCALAS DE WECHSLER

Las escalas de inteligencia preparadas por David Wechsler incluyen tres variaciones sucesivas de aplicación individual, una para adultos, una para niños escolares y otra para preescolares. Además de su uso como medidas de inteligencia general, se ha investigado su posible participación como auxiliar en el diagnóstico psiquiátrico. Par-

tiendo de la observación de que el daño cerebral, el deterioro psicótico y los problemas emocionales pueden afectar algunas funciones intelectuales más que otras, Wechsler y otros psicólogos clínicos argumentaron que el análisis de la ejecución relativa del individuo en diferentes subtests revelaría trastornos psiquiátricos particulares. En el capítulo 17 estudiaremos los problemas y resultados pertinentes al análisis del perfil de las escalas de Wechsler para ejemplificar el uso clínico de las pruebas.

El interés generado por estas escalas y el grado de uso son atestiguados por miles de publicaciones que a la fecha han aparecido sobre ellas. Además de las revisiones usuales publicadas en el *Anuario de Medición Mental*, la investigación relativa a las escalas de Wechsler se revisa periódicamente en las revistas especializadas (Guertin, Frank y Rabin, 1956; Guertin, Ladd, Frank, Rabin y Hiester, 1966; Guertin, Ladd, Frank, Rabin y Hiester, 1971; Guertin, Rabin, Frank y Ladd, 1962; T. D. Hill, Reddon y Jackson, 1985; Littell, 1960; Rabin y Guertin, 1951; I. L. Zimmerman y Woo-Sam, 1972) y ha sido resumida en varios libros (por ejemplo, Forster y Matarazzo, 1990; Gyurke, 1991; Kamphaus, 1993; Kaufman, 1979, 1990, 1994; Sattler, 1988, 1992).

Antecedentes y evolución de las escalas de inteligencia de Wechsler. En 1939 se publicó la primera forma de las escalas de Wechsler, conocida como la Escala de Inteligencia Wechsler-Bellevue, que pretendía ser un test de inteligencia adecuado para adultos. Al presentar por primera vez la escala, Wechsler (1939) señalaba que las pruebas anteriores se habían diseñado ante todo para escolares y que luego se habían adaptado para aplicarse a adultos agregando simplemente reactivos más difíciles de la misma clase, por lo que su contenido solía ser de poco interés para ellos. Indicaba también la dificultad de establecer el *rapport* con adultos a menos que los reactivos tengan un mínimo de validez de facie, de la que carecen muchos reactivos de los tests de inteligencia escritos para hacer referencia a las actividades cotidianas de los escolares.

El énfasis excesivo que la mayor parte de las pruebas pone en la velocidad tiende a perjudicar a las personas mayores. De modo similar, Wechsler creía que la manipulación relativamente rutinaria de las palabras recibía un peso excesivo en los tests tradicionales de inteligencia; además, llamó la atención el hecho de que las normas de edad mental no se aplican a los adultos y señaló que se incluyeron muy pocos adultos en las muestras de estandarización de los tests individuales de inteligencia.

La prueba original de Wechsler-Bellevue estaba destinada a proporcionar respuesta a esas objeciones. En forma y contenido, la escala estableció el patrón básico para las subsecuentes escalas de inteligencia de Wechsler, cada una de las cuales agregó a su vez ciertas mejoras a su predecesora. En 1949 se preparó la Escala de Inteligencia para Niños de Wechsler (WISC) como extensión descendente de la Wechsler-Bellevue (Seashore, Wesman y Doppelt, 1950). Muchos de sus reactivos se tomaron directamente de la prueba para adultos y en cada subtest se agregaron reactivos más sencillos del mismo tipo. El propio Wechsler-Bellevue fue sustituido en 1955 por la Escala de Inteligencia para Adultos de Wechsler (WAIS), que corregía algunas de las deficiencias técnicas de la escala anterior en relación con el tamaño y la representatividad de la muestra normativa y la confiabilidad de los subtests. La más pequeña de la serie de Wechsler, publicada en 1967, es la Escala de Inteligencia para Edades Preescolar y Primaria de Wechsler (WPPSI), diseñada originalmente para edades de cuatro a seis años y medio como extensión descendente del WISC, que a su vez se diseñó para edades de cinco a 15 años.

El desarrollo del WISC fue algo paradójico, ya que Wechsler emprendió la empresa original debido en parte a la necesidad de contar con una escala para adultos que *no* fuera una mera extensión hacia arriba de las escalas para niños. De hecho, la primera edición del WISC fue criticada porque su contenido no estaba lo suficientemente orientado hacia los niños. En la edición revisada (WISC-R), publicada en 1974 y diseñada para niños de seis a 16 años, se hicieron esfuerzos especiales por reemplazar o modificar aquellos reactivos orientados a los adultos para que su contenido se acercara más a las experiencias infantiles comunes; por ejemplo, en el subtest de aritmética se cambiaron los "cigarros" por "caramelos". Otros cambios incluyeron la eliminación de reactivos que podrían ser diferencialmente similares para grupos particulares de niños, así como la inclusión de más mujeres y personas negras en el contenido pictórico de los subtests. Algunos de los subtests incrementaron su longitud para aumentar su confiabilidad y además se introdujeron mejoras en los procedimientos de aplicación y calificación.

Descripción de las escalas. Hasta ahora, cada una de las tres escalas de Wechsler ha pasado por una o más revisiones. Las versiones actuales, publicadas bajo el nombre de David Wechsler incluso después de su muerte en 1981, son la Escala de Inteligencia para Adultos de Wechsler, Revisada (WAIS-R—Wechsler, 1981), que cubre el intervalo de los 16 a los 74 años; la Escala de Inteligencia para Niños de Wechsler, tercera edición (WISC-III—Wechsler, 1991) diseñada para niños de seis a 16 años y 11 meses; y la Escala de Inteligencia para Edades Preescolar y Primaria de Wechsler, Revisada (WPPSI-R—Wechsler, 1989), que ahora cubre el rango de tres a siete años y tres meses. Para 1997 se anticipa la publicación de la tercera edición del WAIS, que está en preparación desde 1992.

El WAIS-R, el WISC-III y el WPPSI-R comparten muchas características, incluyendo su organización básica en las escalas verbal y de ejecución, cada una de las cuales consta de un mínimo de cinco subtests (y un máximo de siete) y proporciona CI de desviación separados. Las calificaciones individuales en los 10 subtests regularmente aplicados (11 en el caso del WAIS-R) se combinan en una Escala de CI que tiene una media de 100 y una *DE* igual a 15, como el CI Verbal y de ejecución. De los 17 tipos diferentes de subtests empleados en el WAIS-R, el WISC-III y el WPPSI-R, ocho (cinco subtests verbales y tres de ejecución) son comunes a las tres escalas. Al aplicar las escalas, se alternan los subtests verbal y de ejecución y se presentan en una secuencia determinada que varía con cada escala.

En las tres escalas, el primer subtest que se aplica es el verbal, lo que ayuda a establecer un buen *rapport*. Se han hecho esfuerzos por evitar el conocimiento especializado. Los primeros reactivos son lo suficientemente sencillos como para ser aprobados por la gran mayoría de los examinados, a menos que sean mentalmente retardados o que tengan problemas de orientación de la realidad, en cuyo caso el examinador puede decidir descontinuar el examen. Las preguntas sobre información en las versiones WAIS-R y WISC-III cubren hechos que la mayor parte de los estadounidenses ha tenido oportunidad de aprender, como "¿Qué mes viene antes de diciembre?" o "¿Quién era Mark Twain?" El WPPSI-R incluye preguntas similares, aunque de menor nivel de dificultad. En realidad inicia con unos pocos reactivos presentados en forma pictórica

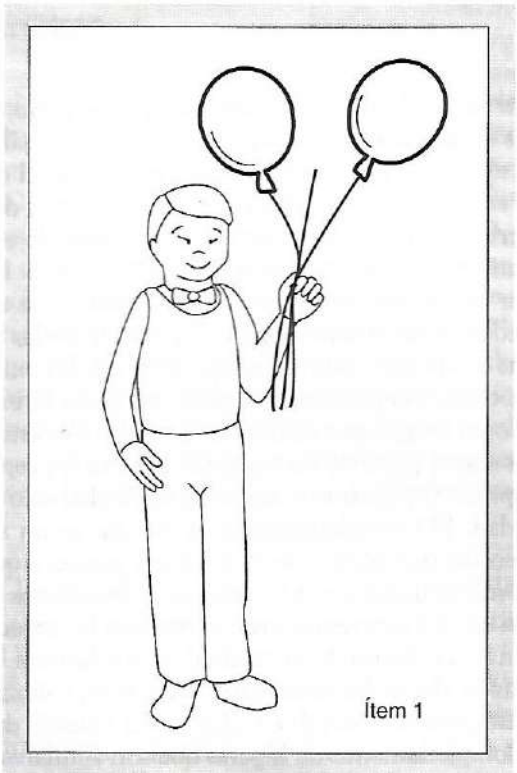
que sólo requieren que se señale la respuesta correcta; por ejemplo, al presentar una imagen de varios objetos domésticos puede preguntarse al niño cuál es el que se emplea para la limpieza. El subtest de aritmética es otra medida verbal que ilustra el amplio grado de dificultad en las escalas de Wechsler. Los reactivos de aritmética más sencillos del WPPSI-R requieren que se señale el objeto ilustrado en un arreglo que ilustra un concepto cuantitativo (como "el más pequeño" o "más"). Los reactivos más complejos pueden incluir tareas como contar o resolver problemas aritméticos, y los más difíciles de éstos pueden requerir una buena comprensión de las fracciones.

En general, los subtests de ejecución de las escalas de Wechsler exigen la manipulación de diversos objetos, como rompecabezas y cubos, así como la exploración visual de materiales impresos, como imágenes o símbolos. Todos los subtests establecen límites de tiempo, pero en la mayor parte de los casos también se le proporcionan puntos de bonificación por la rapidez. En contraste, en la Escala Verbal sólo un subtest (el de aritmética) es de velocidad. El Completamiento de Figuras es un subtest de ejecución, que se encuentra en las tres escalas de Wechsler y requiere que el examinado identifique la parte importante que falta en las imágenes de objetos o escenas comunes. Los reactivos para las edades más tempranas se basan en la inspección visual básica, por ejemplo al presentar el dibujo de un animal al que le falta un miembro; los reactivos más difíciles exigen algo de razonamiento deductivo, conocimiento específico o ambos para identificar qué es lo que falta. La figura 8.3 muestra dos reactivos relativamente sencillos de Completamiento de Figuras que son similares a los empleados en las escalas de Wechsler.

Escalas abreviadas. Desde la publicación de la prueba original de Wechsler-Bellevue se ha formulado una gran cantidad de *escalas abreviadas* o *formas cortas* de las escalas Wechsler con el objeto de reducir sustancialmente el tiempo de aplicación y hacer una estimación del CI de la Escala Completa que pueda evaluarse en términos de las normas publicadas. Una manera de elaborarlas consiste en omitir algunos de los subtests y prorratear las calificaciones, o bien en disminuir el número de reactivos de los subtests.

Su uso y desarrollo para propósitos de identificación rápida se ha visto estimulado por el hecho de que diversas combinaciones de subtests muestran una correlación superior a .90 con los CI de la Escala Completa. Se ha llevado a cabo una extensa investigación (principalmente en poblaciones especiales, como pacientes psiquiátricos y personas con retardo mental) para identificar las combinaciones de dos, tres, cuatro y cinco subtests que predicen de manera más efectiva las puntuaciones obtenidas en la parte Verbal, de Ejecución y los CI de la Escala Completa (Matarazzo, 1972; McCusker, 1994; Sattler, 1988, 1992).

Aunque se ha invertido mucha energía en la integración y supervisión de las formas abreviadas de las escalas de Wechsler, se cuestiona la calidad de los procedimientos empleados para derivar las escalas abreviadas (Silverstein, 1990); por ejemplo, quizá no siempre se justifique el uso de las normas de la Escala Completa original para prorratear las calificaciones totales de las escalas abreviadas. Más aún, al utilizar escalas abreviadas se pierden muchas de las observaciones cualitativas importantes que la aplicación de una escala individual hace posible, por lo que es probable que no sea aconsejable el uso de esas versiones salvo como dispositivos de identificación aproximada.



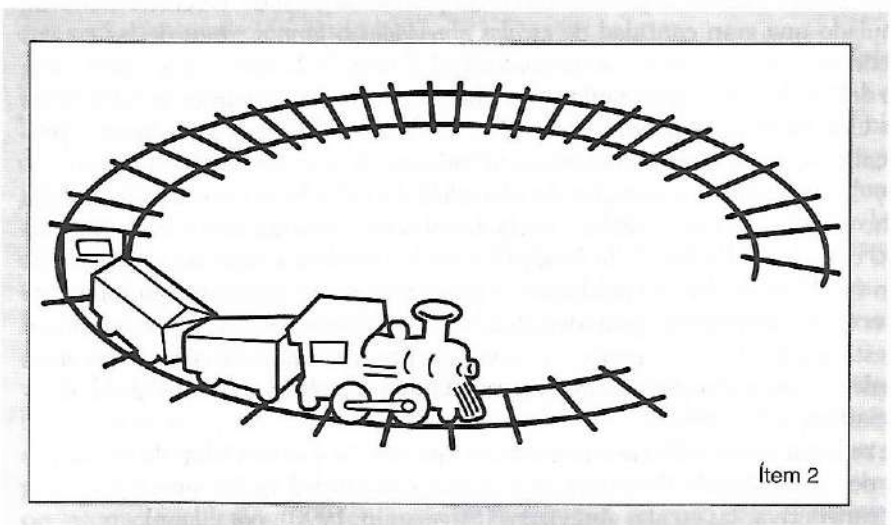


Figura 8.3. Dos reactivos de *Completamiento de Figuras* similares a las que se emplean en la Escala de Inteligencia para Edades Preescolar y Primaria de Wechsler, Revisada. (Cortesía de The Psychological Corporation.)

Normas y calificación. Las muestras de estandarización de las últimas escalas de Wechsler se eligieron con especial cuidado para asegurar su representatividad. Las muestras normativas incluyeron para cada escala aproximadamente 2 000 casos, con igual número de mujeres y hombres distribuidos en los grupos de edad apropiados para cada uno. Los participantes se eligieron para igualar tanto como fuera posible las proporciones de los informes más recientes del censo estadounidense en el momento de la estandarización en relación con variables como región geográfica, raza o etnicidad, nivel ocupacional y escolaridad. En el caso de los niños, se empleó el nivel ocupacional de los padres. En cada revisión las variables empleadas en el plan del muestreo estratificado han cambiado un poco en una dirección que hace a las muestras más inclusivas; por ejemplo, la categoría de etnicidad del WISC-III incluye cuatro grupos (blancos, negros, hispanos y otros) mientras que la versión anterior, el WISC-R, sólo empleaba dos (blancos y no blancos). Además, a diferencia de las escalas anteriores, la muestra del WISC-III incluyó explícitamente cantidades representativas de estudiantes que recibían servicios especiales en ambientes escolares, como es el caso de los niños con problemas de aprendizaje y los de programas para sobredotados.

La popularidad de las escalas de Wechsler, que actualmente son los tests individuales de inteligencia de mayor uso, ha dado lugar a la realización de investigaciones diseñadas para aumentar su utilidad; por ejemplo, como parte de una serie de estudios normativos realizados con ancianos estadounidenses en la Clínica Mayo, se examinó a una muestra de 222 personas de 56 a 97 años con el fin de obtener datos normativos para el WAIS-R que se extendiera al grupo de mayor edad en la muestra de estandarización (Ivnik *et al.*, 1992). Otro estudio obtuvo normas para el WAIS-R basadas en 130 personas de más de 75 años (Ryan, Paolo y Brungardt, 1990).

Las puntuaciones crudas o directas de cada uno de los subtests del Wechsler se transformaron en calificaciones estándares con una media de 10 y una *DE* igual a tres, lo que permite expresar en unidades comparables las calificaciones escaladas de todos los subtests. Las calificaciones de los subtests Verbal, de Ejecución y de la Escala Completa se sumaron y se convirtieron en calificaciones estándares (conocidas como "CI de desviación") con una media de 100 y una *DE* igual a 15. El WISC-III proporciona además índices de calificación para cuatro factores: Comprensión Verbal (ICV), Organización Perceptual (IOP), Libertad de Distracción (ILD) y Velocidad de Procesamiento (IVP). La composición de esos índices de calificación se asemeja a los factores que muchos investigadores independientes derivaron de la versión anterior, WISC-R. Los nuevos índices de calificación se basan en combinaciones de dos o cuatro subtests y tienen una media de 100 y una *DE* igual a 15. Cada una de las tres escalas de Wechsler proporciona además la información necesaria para evaluar la ejecución del sujeto en cada subtest y grupo de subtests en términos de las normas de edad apropiadas.

Confiabilidad. Las escalas de Wechsler proporcionan para cada grupo de edad en cada calificación del subtest coeficientes de confiabilidad de división por mitades,[7] índices de calificación y CI. En todas las escalas la confiabilidad de división por mitades

[7] Salvo aquellos para los cuales es inapropiada la confiabilidad de división de mitades (es decir, Símbolos Dígitos, Codificación, Clavijas De Animales, Búsqueda de Símbolos y Retención de Dígitos).

de los CI de la Escala Completa fluctuaba entre .90 y .98, entre .86 y .97 para los CI Verbales, y entre .85 y .94 para los CI de Ejecución. La confiabilidad de división por mitades de los índices de calificación basados en factores del WISC-III fluctuaron entre .80 y .95. Como era de esperar, la confiabilidad de los subtests de las tres escalas de Wechsler tendía a ser baja, entre .52 y .96, aunque la gran mayoría se encontraba por encima de .70. Es particularmente importante considerar la confiabilidad de los subtests al evaluar la significancia de las diferencias entre las calificaciones obtenidas por los mismos individuos, como sucede en el análisis del perfil (J. H. Kramer, 1990, 1993; Sattler, 1988, 1992). Los manuales de Wechsler también presentan los errores de medición para todas sus calificaciones. Para los CI Verbales los errores varían entre 2.50 y 4.98 puntos, para los CI de ejecución, entre 3.67 y 4.97, y para la escala completa están por debajo de 4.00, todo esto nos permite concluir, por ejemplo, que hay una posibilidad aproximada de 2:1 de que el CI verdadero de un individuo en la escala completa caiga cuatro puntos del CI que obtuvo en dicha escala.

En cada revisión se obtuvieron también con más cuidado los datos sobre la confiabilidad del retest (o estabilidad) de las calificaciones de las escalas de Wechsler. Los coeficientes de estabilidad suelen ser mayores para los adultos que para los niños, e invariablemente se encuentra que en el retest hay incrementos de entre dos y 13 puntos en los diversos CI, con intervalos de 12 días a nueve semanas; por lo general, los CI de la escala completa se elevan entre cinco y siete puntos. Al volver a examinar a los individuos luego de un breve periodo, es necesario tener en cuenta este esperado, aunque ligero, efecto de la práctica.

Los manuales del WPPSI-R y el WISC-III son los más recientes de una serie que con el tiempo han mejorado mucho. Una novedad es que incluyen los cálculos de la confiabilidad del calificador en los subtests cuya calificación requiere de un juicio considerable. Los datos indican que esos subtests pueden ser calificados con una confiabilidad aceptable por personas capacitadas y que han recibido una práctica adecuada. Otra innovación es que proporcionan datos sobre las diferencias de calificación intraindividuales. Además de las tablas que muestran la cantidad de la diferencia que se requiere para la significancia estadística, los manuales presentan las frecuencias de las diferencias encontradas dentro de la muestra de estandarización, información que resulta particularmente valiosa en el uso clínico de estas pruebas (véase el capítulo 17).

Validez. En ninguna parte fue más evidente la mejora de los manuales de Wechsler que en el área de la validez. En 1981, cuando se publicó el WAIS-R, el manual no presentaba más datos sobre la validez del instrumento que un par de estudios de la correlación entre esta versión y las escalas anteriores. La información sobre la validez era incluso menor a la que podía encontrarse en los manuales del WPPSI y el WISC-R, que al menos presentaban datos sobre las correlaciones entre éstas y otras escalas de inteligencia, como las de Stanford-Binet; sin embargo, esta limitación se vio compensada en parte en la investigación independiente sobre la validez de todas las escalas de Wechsler.[8] Hasta cierto punto, el poco énfasis en la información sobre la validez en los

[8] Véase Dean, 1977, 1979, 1980; Gutkin y Reynolds, 1981; G. P. Hollenbeck y Kaufman, 1973; Karnes y Brown, 1980; Kaufman, 1975; Kaufman y Hollenbeck, 1974; Leckliter, Matarazzo y Silverstein, 1986; Silverstein, 1982a, 1982b; y Waller y Waldman, 1990.

manuales de Wechsler obedece a su convicción de que las tareas de sus escalas cubren una gama de habilidades específicas que es suficiente para asegurar la validez en la evaluación de la inteligencia general (Wechsler, 1958; Zachary, 1990).

El tratamiento que hizo Wechsler de la validez refleja en esencia una orientación a la descripción del contenido, aunque tenía visos de la aproximación de la identificación del constructo, con pocos datos de apoyo; por ejemplo, las correlaciones entre las escalas de Wechsler y otras medidas globales de inteligencia, como el Stanford-Binet, se agrupaban alrededor de .80. Además, los análisis factoriales realizados por investigadores independientes arrojaron resultados consistentes. Casi invariablemente se encontró en todos los grupos de edad un factor verbal y uno perceptual-organizacional o no verbal; asimismo, en las escalas para niños mayores y adultos por lo general aparecían factores adicionales de memoria o atención. El uso de las escalas de Wechsler para tomar decisiones de selección ocupacional y educativa resultó justificada, hasta cierto punto, por datos que mostraban diferencias en la dirección esperada entre varios grupos.

En contraste, en los últimos instrumentos de Wechsler abundan los datos sobre todos los tipos de validez. Los manuales del WPPSI-R y el WISC-III dedican respectivamente 21 y 38 páginas a la revisión de la validez que contrastan con las dos páginas que dedicaba al tema el manual del WAIS-R. Los datos sobre la validación referida al constructo se encuentran en las intercorrelaciones entre los subtests y los análisis factoriales de las puntuaciones. En los distintos grupos de edad de las muestras de estandarización se encontraron correlaciones promedio entre las escalas Verbal y de Ejecución de .74 para el WAIS-R, de .66 para el WISC-III y de .59 para el WPPSI-R. Estas cifras indican la presencia de un factor general que es corroborada por la mayor parte de los análisis factoriales de las escalas.

Los análisis factoriales de los nueve grupos de edad de la muestra de estandarización del WAIS-R indicaron repetidamente que un modelo de tres factores da una mejor explicación de las correlaciones entre los 11 subtests. Esos factores, que parecen ser generalizables a diferentes muestras, son Comprensión Verbal, Organización Perceptual y Memoria/Libertad de Distracción (Leckliter *et al.*, 1986; Waller y Waldman, 1990). Al aplicar el mismo tipo de análisis a los datos de estandarización del WPPSI-R descritos en el manual se obtiene una solución de dos factores que es consistente con la organización de los subtests en las escalas Verbal y de Ejecución (Blaha y Wallbrown, 1991; LoBello y Gulgoz, 1991; B. J. Stone, Gridley y Gyurke, 1991). Por otro lado, los datos del WISC-III, que desde el inicio se sometieron a los análisis exploratorios y confirmatorios descritos en el manual, parecen ser mejor explicados por un modelo de cuatro factores (Comprensión Verbal, Organización Perceptual, Libertad de Distracción y Velocidad de Procesamiento) incorporados a la calificación estándar del WISC-III.

Los manuales del WPPSI-R y del WISC-III también presentan datos de validez de numerosos estudios, aunque con muestras pequeñas, que muestran su correlación con otras pruebas de aplicación individual. En el caso del WISC-III, también se presentan algunas correlaciones con las escalas de aprovechamiento administradas grupalmente y con las calificaciones escolares. Se investigó además la utilidad diagnóstica o predictiva del WISC-III y del WPPSI-R en una serie de estudios con grupos especiales, como niños sobredotados, retardados, con problemas de aprendizaje, etcétera.

Observaciones finales sobre las escalas de Wechsler

Las versiones sucesivas de las tres escalas de Wechsler reflejan un nivel creciente de sofisticación y experiencia en la elaboración de pruebas que corresponde a las décadas en las que se desarrollaron. En comparación con otras pruebas de aplicación individual, su principal ventaja proviene del tamaño y representatividad de las muestras de estandarización (particularmente en las poblaciones adulta y preescolar) y la calidad técnica de sus procedimientos en la elaboración de pruebas. El tratamiento de la confiabilidad y la validez en el manual del WISC-III es particularmente encomiable. La popularidad de las escalas de Wechsler les asegura, por el momento, una base de investigación en constante expansión. Los usuarios también disponen de materiales auxiliares, como programas de interpretación asistidos por computadora, manuales para el entrenamiento de examinadores (por ejemplo, Fantuzzo, Blakey y Gorsuch, 1989) y guías de interpretación (por ejemplo, Kaufman, 1994; Nicholson y Alcorn, 1994; Whitworth y Sutton, 1993). Sin embargo, algunos críticos han advertido que incluso las más recientes y mejoradas versiones de las escalas de Wechsler pueden resultar obsoletas pronto dada la demanda de vínculos entre los instrumentos de evaluación y las estrategias de intervención (Shaw, Swerdlik y Laurent, 1993; Sternberg, 1993). A este respecto, el punto débil de todas las escalas de Wechsler es que carecen de fundamentos teóricos, lo que hace difícil encontrar una base coherente para la interpretación. Además, la composición de las escalas parece suponer, por similitudes superficiales entre los materiales y tareas de la prueba, que los dominios de habilidad cubiertos por los subtests entre los niveles de edad son los mismos, suposición que no puede justificarse al considerar lo que ahora sabemos sobre los cambios que ocurren durante el ciclo vital en la naturaleza de la inteligencia (véase el capítulo 11).

LAS ESCALAS DE KAUFMAN

Las escalas de Kaufman son instrumentos clínicos de aplicación individual diseñadas para muchos de los usos para los que se desarrollaron pruebas como las escalas de Stanford-Binet y las de Wechsler (Kaufman y Kaufman, 1983a, 1983b, 1990, 1993). Preparadas en la década de los ochenta y a inicio de los noventa, las escalas de Kaufman incorporan los avances recientes en la elaboración de pruebas. En particular, la Batería de Evaluación para Niños de Kaufman, K-ABC (Kaufman y Kaufman, 1983a, 1983b) y el Test de Inteligencia para Adolescentes y Adultos de Kaufman, KAIT (Kaufman y Kaufman, 1993) representan un intento por parte de sus autores, ambos participantes en la elaboración del WISC-R, por apartarse de la posición ateórica de las antiguas escalas de inteligencia. Ambos buscaban crear instrumentos que por su diseño pudieran afianzarse en las teorías en evolución sobre la inteligencia, que incluyeran tareas apropiadas para el desarrollo y proporcionaran información útil para diversas situaciones de evaluación.

Batería de evaluación para niños de Kaufman (K-ABC)

Evaluación general. El diseño de la K-ABC empezó con la formulación de los constructos que se pretendía evaluar. De acuerdo con la orientación general de la psico-

logía cognoscitiva, la atención se centró en el procesamiento de información. La aproximación particular elegida para diferenciar entre el Procesamiento Simultáneo, representado por siete subtests y el Procesamiento Secuencial, representado por tres (J. P. Das, 1984; J. P. Das, Kirby y Jarman, 1975, 1979; J. P. Das y Molloy, 1975; Kaufman y Kaufman, 1983b, capítulo dos; Luria, 1966). Los subtests de Procesamiento Simultáneo requieren la síntesis y organización del contenido espacial y visual-perceptual que puede revisarse en forma global. Los subtests de Procesamiento Secuencial requieren del ordenamiento serial o temporal, utilizan contenido verbal, numérico, visual-perceptual y la memoria a corto plazo. Algunas de las tareas presentadas en la Escala de Procesamiento Mental son similares a las que se utilizan en los exámenes neuropsicológicos (que revisaremos en el capítulo 17) y ésta es justamente la razón por la que se eligieron.

La batería también incluye una Escala de Aprovechamiento compuesta por seis subtests, que aunque evalúan la habilidad en lectura, aritmética, léxico e información general, se diseñaron explícitamente para *no* medir el conocimiento verdadero que se enseña en la escuela, por lo que son más parecidas a las tareas incluidas en las pruebas tradicionales de aptitud o de inteligencia que a las que se presentan en las pruebas de aprovechamiento escolar; por ejemplo, en la prueba de aritmética se le presenta al niño una serie de imágenes de una visita familiar al zoológico y éste responde contando y realizando operaciones numéricas simples con los objetos en cada lámina; por su parte la lectura de comprensión se demuestra realizando las acciones descritas en cada enunciado que el niño lee.

La K-ABC se estandarizó en una muestra nacional estadounidense de 2 000 niños (con edades de dos y medio a 12 y medio). Además, se examinó a niños blancos y negros para desarrollar normas socioculturales de raza y educación de los padres que pudieran auxiliar en la interpretación. El diseño y organización de la K-ABC también pretendía responder a las necesidades de examinación de grupos especiales como los niños discapacitados, los de minorías culturales y lingüísticas, así como para ayudar en el diagnóstico de problemas de aprendizaje (Kamphaus, Kaufman y Harrison, 1990). La batería proporciona cuatro calificaciones globales: Procesamiento Secuencial, Procesamiento Simultáneo, Procesamiento Mental Compuesto (que combina a las dos anteriores) y Aprovechamiento, cada una de las cuales es una calificación estándar con una media de 100 y una *DE* de 15.

Evaluación general. La K-ABC tiene muchos rasgos positivos, tanto técnicos como prácticos,[9] que en principio permiten frenar la tendencia popular a etiquetar a los niños con un único número evaluativo, como el CI, gracias al uso de calificaciones múltiples, análisis de perfiles e interpretaciones diagnósticas, especialmente como las que se explican en los capítulos 5 y 6 del *Manual interpretativo* (Kaufman y Kaufman, 1983b). El capítulo 6 del manual proporciona un ejemplo excelente del ciclo de generación y comprobación de hipótesis que es la esencia de la aproximación clínica al diagnóstico. Más aún, en su esfuerzo por desvanecer algunas creencias erróneas comunes los autores establecen con claridad en el capítulo inicial del *Manual interpretativo*

[9] Para revisiones y análisis críticos véase especialmente T. L. Miller (1984), así como Anastasi (1984a, 1985c), Coffman (1985), Kamphaus (1990), Kline, Snyder y Castellanos (1996), y Page (1985).

(Kaufman y Kaufman, 1983b, pp. 20–24) que la K-ABC *no es* "una medida de habilidades innatas o inmutables", y agregan que "todas las tareas cognoscitivas se consideran como medidas de lo que el individuo ha aprendido". Previenen, adecuadamente, que al igual que otra prueba, la K-ABC no es "la batería de pruebas completa", que para satisfacer necesidades individuales debe ser complementada por otros instrumentos.

Sin embargo, a pesar de las advertencias de los autores, el uso del término "pruebas de aprovechamiento" puede haber sido una elección desafortunada por las creencias erróneas sobre la relación entre las pruebas de aptitud y aprovechamiento. Una prueba puede llamarse sin duda de aprovechamiento cuando está muy vinculada a un contenido instruccional identificable específico, al que se supone que han sido expuestos los examinados. Lo anterior no se cumple en las pruebas llamadas de "aprovechamiento" en la K-ABC, ya que en éstas se hicieron esfuerzos especiales por desvincularlas de la información específica adquirida en el aula. De hecho, dentro del continuo de habilidades desarrolladas, estas pruebas están más cercanas a la aptitud que al aprovechamiento, conclusión que corroboran las correlaciones entre los subtests. En consecuencia, no se justifica el uso de una terminología que ha adquirido un significado excesivo, tiene implicaciones no pretendidas y fomenta las creencias populares erróneas.

La aseveración explícita de un marco teórico como guía para la especificación de las tareas y el desarrollo de reactivos en la K-ABC fue una bienvenida innovación que es congruente con los principios de elaboración de buenas pruebas. No obstante, más de una década después de su aparición, persisten interrogantes al respecto de si la orientación teórica que se eligió fue la mejor para el propósito perseguido. En concreto, se ha cuestionado la distinción entre procesamiento simultáneo y secuencial como base para comprender la ejecución en la K-ABC y se ha argumentado que los dos conjuntos de subtests así designados también podrían caracterizarse como pruebas de razonamiento verbal y no verbal (J. P. Das, 1984; Goetz y Hall, 1984; A. R Jensen, 1984; Keith, 1985; Keith y Dunbar, 1984).

Por otro lado, la investigación ya acumulada sobre la K-ABC postula que sus calificaciones globales son similares a las del WISC-R en términos de su validez predictiva y el grado en que miden la "inteligencia general" (Kamphaus, 1990). Por su menor dependencia de las capacidades verbales, la K-ABC puede ser el instrumento a elegir para niños con un dominio insuficiente del inglés o con deficiencias auditivas. En *Clinical and Research Applications of the K-ABC* (Aplicaciones clínicas y de investigación de la K-ABC) (Kamphaus y Reynolds, 1987, capítulo 8) puede encontrarse una presentación equilibrada de las ventajas y las limitaciones de este instrumento relativamente nuevo.

Test de inteligencia para adolescentes y adultos de Kaufman (KAIT)

Naturaleza y desarrollo. El KAIT (Kaufman y Kaufman, 1993) se diseñó como medida de inteligencia para edades de 11 a 85 o mayores. Representa un intento por integrar la teoría de la inteligencia fluida y cristalizada propuesta por Horn y Cattell (1966) con las ideas que otros teóricos han expuesto sobre la inteligencia adulta (Golden, 1981; Luria, 1980; Piaget, 1972).

La batería se compone de una Escala Cristalizada, que mide conceptos adquiridos como escolaridad y aculturación, y una Escala Fluida que abarca la habilidad para re-

solver nuevos problemas, también se cuenta con una Batería Central o Medular que consta de tres subtest en cada escala; y si se agrega cualquiera de los cuatro subtests específicos, puede emplearse esta Batería Ampliada en individuos de los que se sospecha daño neurológico. El KAIT incluye además un breve examen del Estado Mental para evaluar la atención y orientación en los examinados con un daño cognoscitivo demasiado grande para enfrentar la aplicación de la batería completa.

Evaluación general. En términos de sus cualidades técnicas, el KAIT parece cumplir los estándares psicométricos tan bien como cualquiera de los principales tests de inteligencia de la generación actual. Su muestra normativa es adecuada y la confiabilidad y datos de validez presentados en el manual parecen prometedores. La aplicación de la batería es relativamente sencilla, y el manual incluye información que es de gran ayuda en la aplicación y calificación del instrumento (por ejemplo, nos dice qué hacer cuando alguien responde en un idioma extranjero).

No obstante, lo que realmente distingue al KAIT de otras escalas de inteligencia para adultos es el cuidado que se tuvo para formar y probar su banco original de más de 2 500 reactivos diseñados para atraer a los examinados adultos. En general requieren la clase de solución de problema que caracteriza al pensamiento operacional formal de Piaget y las funciones de planeación evaluativa que según Luria (1980) y Golden (1981) caracterizan al pensamiento adulto. Como resultado, en su mayor parte son desusuales e interesantes. Muchos de los subtests parecen un juego incluso en sus títulos, como es el caso de Caras Famosas, Códigos Misteriosos y Dobles Significados. Otras presentan tareas novedosas, como el Aprendizaje de Jeroglíficos, en el que los examinados aprenden las palabras asociadas con un jeroglífico (dibujo) particular y luego "leen" frases o enunciados compuestos por jeroglíficos. En la figura 8.4 se muestra un ejemplo. La prueba crucial para el KAIT, como para cualquier nuevo instrumento, estriba en si su atractivo será suficiente para generar el volumen de investigación y aplicaciones que le permitan componer una base de datos rica y duradera.

Test Breve de Inteligencia de Kaufman (K-BIT)

El Test Breve de Inteligencia de Kaufman, K-BIT (Kaufman y Kaufman, 1990) se diseñó como instrumento para la estimación rápida del nivel de funcionamiento intelectual. Aunque se aplica de manera individual, la prueba es sencilla y puede ser aplicada por un técnico especializado. El K-BIT cubre el intervalo de edad de cuatro a 90 años; sus normas se establecieron concurrentemente con las del KAIT utilizando alrededor del 20 por ciento de la muestra de estandarización del KAIT compuesta por 2 000 personas.

La K-BIT no es una versión abreviada de la K-ABC ni del KAIT. Consta de un subtest verbal compuesto por 45 reactivos de Vocabulario Expresivo y 37 definiciones, y un subtest no verbal constituido por 48 matrices. Como en el caso de las otras escalas de Kaufman, las tres calificaciones (verbal, no verbal y compuesta) proporcionadas por el K-BIT se expresan en unidades de CI de desviación; además, dada la extensión de los subtests, sus coeficientes de confiabilidad son mayores que en otras escalas. No obstante, en relación con el tema crucial de la correlación entre sus calificaciones y las de las escalas completas, el K-BIT no parece representar un avance sobre las formas cortas de otros tests de inteligencia.

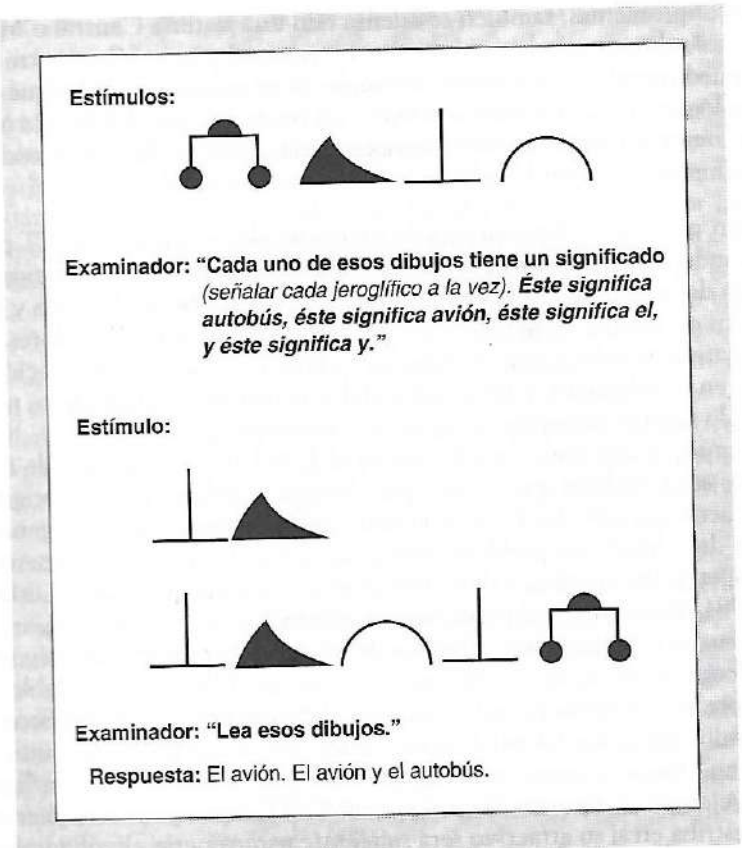


Figura 8.4. Ejemplo del subtest de Aprendizaje de Jeroglíficos de la Escala de Inteligencia para Adolescentes y Adultos de Kaufman.

(De Kaufman y Kaufman, 1993, p. 5. Copyright © 1993 de la American Guidance Service, Inc. Reproducido con permiso del editor.)

ESCALAS DE HABILIDAD DIFERENCIAL

Las Escalas de Habilidad Diferencial (*Differential Ability Scales*, DAS) de C. D. Elliott (1990a, 1990b) constituyen una revisión y extensión de las Escalas Británicas de Habilidad (*British Ability Scales*, BAS) elaboradas en Inglaterra en la década de los setenta (Elliott, Murray y Pearson, 1979). La DAS comparte con las versiones actuales de las escalas de Stanford-Binet y de Wechsler el objetivo de clasificar a las personas en términos de su nivel general de habilidad y producir perfiles tanto de sus fortalezas y puntos débiles intelectuales; sin embargo, tanto en sus procedimientos como en sus características técnicas se distingue de otras escalas por el hecho de que incorpora muchos avances de la teoría y práctica psicométrica. A este respecto, es importante advertir que el prefacio del manual de la DAS afirma que los términos "inteligencia" y "CI" no forman parte de su vocabulario (Elliott, 1990a, p. vi), y que la organización,

calificación e interpretación de la escala se dirigen a las conductas definidas con precisión que están siendo evaluadas. Esta declaración explícita, que aparece por primera vez en el manual de una escala de habilidad general, debería ayudar a disipar los estereotipos e ideas erróneas asociadas con el uso popular de esos términos.

Descripción. La DAS se planeó principalmente para medir habilidades específicas con una confiabilidad adecuada y para auxiliar en metas más complejas de la evaluación individual, es decir, el diagnóstico diferencial y la planeación del tratamiento. La elección de las tareas incluidas en la batería fue orientada por bases teóricas y empíricas. El fundamento teórico es ecléctico y flexible; utiliza una aproximación jerárquica de las habilidades mentales que dan cabida a diferentes niveles de generalidad y proporcionan una amplia base de información de la que pueden derivarse hipótesis sobre los examinados. Esta estructura es comparable con los descubrimientos empíricos sobre el desarrollo de las habilidades cognoscitivas. A diferencia de las escalas anteriores que forzaban los datos en un modelo teórico independientemente del ajuste, la DAS sólo conserva aquellos componentes en que coincidían las bases empíricas y teóricas.

Como se muestra en la figura 8.5, la DAS consta de 20 subtests organizados en tres componentes principales: (1) los subtests centrales, (2) los de diagnóstico y (3) los de aprovechamiento. Los títulos de estos tests y subtests describen las tareas que incluyen y en gran medida se explican por sí mismos. Doce subtests centrales y cinco de diagnóstico constituyen la *batería cognoscitiva*, que a su vez se divide en niveles preescolar y escolar. Para nivel preescolar se dispone de cuatro subtests centrales para las edades de dos años con seis meses a tres años con cinco meses, y seis para las edades de tres años con seis meses a cinco años con 11 meses,[10] y con seis para el nivel escolar (edades de seis a 17 años con 11 meses). Para cada nivel de edad, se utiliza la suma de los subtests centrales para obtener la calificación de *Habilidad Conceptual General* (HCG), que en el caso de la DAS es la calificación resumida global. Las pruebas diseñadas como "centrales" de la batería DAS son las que tienen mayores cargas del factor general de la batería o "g". Por otro lado, los subtests de diagnóstico tienen bajas correlaciones con g y no se agrupan en factores de grupo, lo que significa que miden habilidades relativamente independientes; cuando las pruebas de diagnóstico son adecuadas, se dispone de dos a cinco según el nivel de edad. Por último, se dispone de tres pruebas de aprovechamiento que en general se aplican después de los seis años de edad.

Como puede verse en la figura 8.5, algunos subtests de cada uno de los principales componentes de la DAS pueden aplicarse e interpretarse de manera apropiada fuera del nivel de edad para el que normalmente se recomiendan. La ejecución en los subtests designados como "rango extendido de edad" y "fuera de nivel" puede compararse con las normas de las muestras de edad apropiadas obtenidas durante el proceso de estandarización de la DAS. Los subtests en la categoría del *rango extendido de edad* pueden usarse como medidas adicionales de diagnóstico cuando su contenido sea pertinente para la evaluación de un individuo determinado; por ejemplo, el subtest de Construcción con Cubos puede aplicarse a niños de tres años con seis meses a cuatro años con 11 meses

[10] Al informar edades es frecuente emplear dos puntos para separar años y meses; así, "2:6" significa "dos años con seis meses". Esta notación se empleará en las figuras del libro siempre que se presente una serie de edades pero no en el cuerpo del texto.

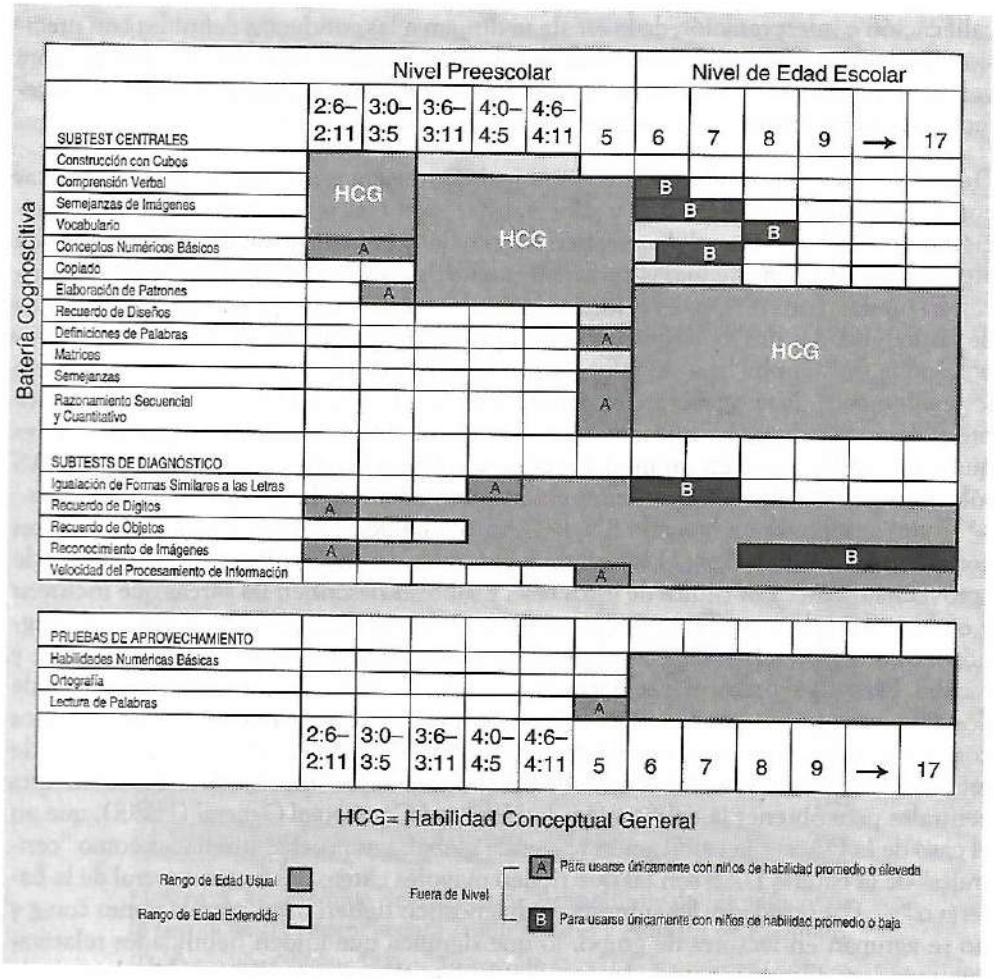


Figura 8.5. Organización de las Escalas de Habilidad Diferencial.

(Adaptado de Elliott, 1990b, p. 4. Copyright © 1990 por The Psychological Corporation. Reproducido con autorización del editor.)

para obtener más información sobre las capacidades perceptuales y motoras finas de lo que es posible con la batería central para esa edad. Por otro lado, los subtests normados para la examinación *fuera de nivel* (señalados con una "A" o "B" en la figura 8.5) sólo son apropiados para individuos cuyo nivel de habilidad es de promedio alto ("A") o de promedio bajo ("B"). Una ventaja de esta característica es que los examinadores pueden evaluar con un grado de exactitud sin precedentes las habilidades de individuos excepcionales que funcionan a niveles desusualmente elevados o bajos para su edad.

Escalas y normas. Una razón importante detrás de las ventajas conceptuales y técnicas de la DAS es que se benefició del razonamiento y la investigación realizados

cuando se elaboró su predecesora, la BAS. La planeación, elaboración y estandarización de la BAS se llevaron a cabo por un periodo aproximado de dos décadas, durante las cuales hubo avances significativos en la teoría y práctica psicométricas. Así, la DAS es un nuevo instrumento con muchos rasgos actualizados, pero además refleja el conocimiento y la experiencia obtenidos desde la década de los sesenta a los ochenta.

La estandarización de la DAS fue ejemplar lo mismo en el tamaño de la muestra que en el cuidado con que se formó. La muestra incluyó 3 475 casos, un número bastante mayor que el empleado en las pruebas de aplicación individual. Se planeó para representar a la población objetivo de todas las personas no internadas en instituciones con dominio del inglés, de dos años con seis meses a 17 años con 11 meses que vivían en los Estados Unidos en el momento de la obtención de los datos (1987–1989). La estratificación se basó principalmente en la edad, género, raza/etnicidad, educación de los padres y región geográfica. Las cifras de la población objetivo no se basaron únicamente en una variable demográfica a la vez, como parece ser el caso, sino en variables combinadas; por ejemplo, la distribución de la educación para mujeres blancas que habitaban en el noreste entre la muestra de la DAS se aproximó a la distribución correspondiente de esas mujeres en la población de los estados del noreste en los Estados Unidos. Las cifras objetivo se calcularon a partir de los datos crudos obtenidos del Ministerio del Censo para el más reciente periodo disponible en ese momento. Aunque la muestra de estandarización fue representativa de la composición racial y étnica de la población (categorizada como negra, hispana, blanca y otras), se obtuvieron protocolos adicionales para alrededor de 300 a 600 negros e hispanos que sólo se emplearon para análisis del sesgo. No se excluyó del grupo normativo a los estudiantes que recibían educación especial, como las clases para niños con deficiencias moderadas o con talentos excepcionales, lo que significa la posibilidad de incluir a toda la población de edad escolar y no únicamente a un grupo "normal".

La DAS utilizó un modelo de un parámetro de la Teoría de Respuesta al Ítem (TRI)[11] para permitir la calibración de los grados de dificultad de cada ítem o reactivo, por lo que es posible examinar a los individuos con los reactivos más adecuados para su nivel de habilidad (es decir, un examen adaptativo). La calificación individual se basa en el número y el grado de dificultad de los reactivos completados. Los datos se colocan en una escala común no normativa, que luego se utiliza para convertir las puntuaciones crudas de cada subtest en calificaciones de habilidad. Para identificar y eliminar los reactivos que no se ajustaban al modelo se utilizaron también estadísticas de la exactitud del ajuste, basadas en la correspondencia entre las respuestas anticipadas y observadas, lo que tuvo el efecto de producir conjuntos de reactivos más homogéneos.

En la aplicación real de la DAS, la estrategia adaptativa de examinación se pone en práctica al designar *puntos de partida* (basados en la edad), *puntos de decisión* (basados en la ejecución desde el punto de partida hasta el punto de decisión) y *reglas alternativas de interrupción*, que son específicas para cada subtest. Los conjuntos de reactivos incluidos entre esos puntos se identificaron empíricamente para producir el equilibrio más eficiente entre confiabilidad y extensión de la prueba. La principal ventaja de esta estrategia adaptativa de examinación de la DAS es la flexibilidad, que permite al examinador seleccionar reactivos de los subtests apropiados para cada exa-

[11] Véase el capítulo 7 para una explicación.

minado. Sin embargo, la posibilidad de obtener estimaciones de la habilidad basadas en una escala común de dificultad del reactivo, incluso cuando se han aplicado diferentes reactivos de los subtests, tiene la ventaja adicional de permitir la comparación de las calificaciones obtenidas por individuos diferentes en el subtest de habilidad y de un único individuo en distintas ocasiones. Esta característica hace a la DAS y a otros instrumentos que siguen las mismas líneas, particularmente adecuados para estudios del desarrollo que utilizan métodos longitudinales y transversales.

Una vez que se han obtenido las puntuaciones de habilidad en los subtests de la batería cognoscitiva, pueden convertirse en calificaciones estándares normalizadas con una media de 50 y una *DE* de 10 (calificaciones T) o en percentiles equivalentes. Ambos tipos de calificaciones se presentan dentro de cada grupo de edad. Las pruebas de aprovechamiento utilizan calificaciones estándares con una media de 100 y una *DE* de 15, en lugar de calificaciones T, y calificaciones por grado en lugar de percentiles basados en la edad. También es posible obtener calificaciones equivalentes a la edad para todos los subtests de la DAS. Para las pruebas de aprovechamiento pueden obtenerse también equivalentes de grado. Tales equivalentes indican la edad (o el grado) en que la calificación de la habilidad del examinado corresponde a la calificación promedio. Como las medidas cognoscitivas y de aprovechamiento de la DAS se desarrollaron y normaron al mismo tiempo, las comparaciones normativas permitidas por todas esas transformaciones hacen posible que el usuario aborde diversas interrogantes apropiadas para la investigación precisa de problemas particulares.

Para derivar la calificación (o calificaciones) compuesta(s) de los diferentes niveles de edad se suman las puntuaciones de los subtests centrales de la DAS y se expresan como calificaciones estándares con una media de 100 y una *DE* de 15. En la figura 8.6 puede observarse que la única calificación compuesta para los niños de menor edad (dos años con seis meses a tres años con cinco meses) es la HCG; para las edades de tres años con seis meses a cinco años con 11 meses la DAS proporciona dos calificaciones agrupadas (Habilidad Verbal y No Verbal) además de la HCG; mientras que para los de edad escolar (de seis a 17 años con 11 meses) pueden derivarse tres agrupamientos (Habilidad Verbal, Razonamiento No Verbal y Habilidad Espacial). Mediante extrapolaciones de la relación entre las calificaciones crudas de ejecución y la HCG a edades diferentes, también es posible extender el uso de las normas HCG a niveles de ejecución por debajo de las normas estándares. Esto se hizo para facilitar la evaluación de individuos con retardo profundo cuya probabilidad de verse representados en la muestra normativa era escasa.

Confiabilidad y validez. Los índices de confiabilidad de la DAS resultan favorables en comparación con los de otros tests de inteligencia. Los coeficientes de confiabilidad test-retest de la HCG y las calificaciones agrupadas, con intervalos de dos a siete semanas entre las pruebas, van de .79 a .94. Para los subtests, los estimados comparables de confiabilidad en el retest van de .38 a .94, aunque la abrumadora mayoría cae entre .60 y .90. La confiabilidad entre calificadores para los subtests con respuestas abiertas cuya calificación[12] requiere de una cantidad sustancial de juicio se agrupan a la mitad de los .90.

[12] Los subtests en cuestión son Definiciones de Palabras, Semejanzas, Copiado y Recuerdo de Diseños.

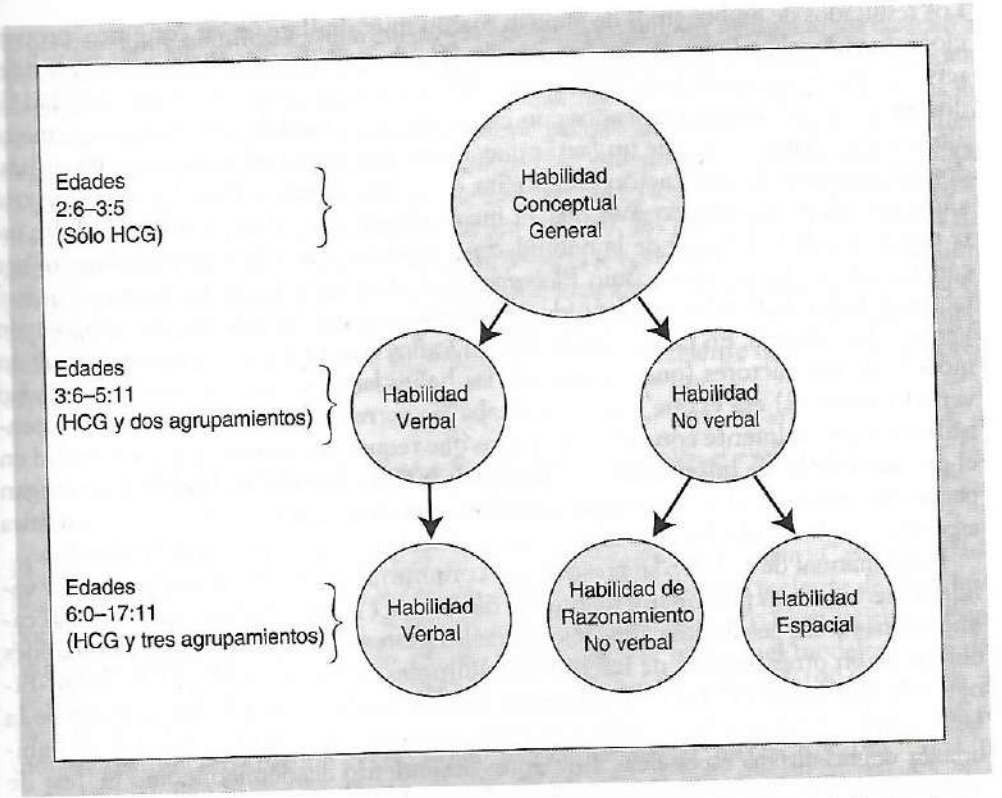


Figura 8.6. Calificaciones compuestas de la Batería Cognoscitiva de la Escala de Habilidad Diferencial.

(Tomado de Elliott, 1990b, p. 21. Copyright © 1990 por The Psychological Corporation. Reproducido con permiso del editor.)

La confiabilidad por consistencia interna de la DAS se evaluó mediante la aplicación de la Teoría de Respuesta al Ítem (TRI), lo que permitió el cálculo preciso de la confiabilidad y los errores de medición correspondientes a cada posible calificación en un subtest. Los resultados, que varían mucho a lo largo del espectro de habilidad, corroboran la bien conocida tendencia a que los coeficientes de confiabilidad sean menores para las personas en los extremos del rango de calificación que para quienes se encuentran cerca del centro. Para la DAS, las estimaciones por confiabilidad de consistencia interna, para todos los niveles apropiados de edad, van de .66 a.95 para los subtests, de .86 a .94 para las calificaciones agrupadas, y de .89 a .96 para las calificaciones HCG. El manual de la DAS también presenta algunas comparaciones entre los coeficientes de consistencia interna tradicionales y los basados en la TRI. En estos casos, se observó un estrecho acuerdo entre las cifras de confiabilidad obtenidas por ambos métodos.

La validez de la DAS ha sido explorada desde dos puntos de vista: interno y externo. En relación con la validez interna, se utilizaron análisis factoriales exploratorio y confirmatorio para derivar la estructura de las calificaciones DAS (véase la figura 8.6).

Los resultados de ambos tipos de análisis fueron muy similares y, en conjunto, proporcionan evidencia adicional que favorece la idea de que al crecer hay una mayor diferenciación de las habilidades (Anastasi, 1970). Para los niños del grupo de edad de dos años con seis meses a tres años con cinco meses, el modelo que muestra un mejor ajuste de los datos es uno de un factor que abarca los cuatro subtests centrales utilizados para generar la calificación HCG para ese grupo de edad. Para los niños de tres años con seis meses a cinco años con 11 meses, surgió una solución bifactorial. En este rango de edad, el factor de la habilidad no verbal se identificó principalmente por cargas altas en los subtests como *Elaboración de Patrones* y *Copiado*, mientras que el factor de habilidad verbal era definido principalmente por los subtests de *Comprensión Verbal* y *Vocabulario*. En las edades de seis a 17 años con 11 meses, se encontró que un modelo de tres factores (que comprende las habilidades verbal, de razonamiento no verbal y espacial) era el que mejor explicaba las correlaciones entre los subtests centrales que generalmente constaban de tareas que requerían de memoria y velocidad en el procesamiento de información. El hecho de que los subtests de diagnóstico tengan pequeñas cargas en el factor común (o "g") y una cantidad considerable de varianza específica los hace ideales para identificar las fortalezas y debilidades del individuo.

En el manual de la DAS se presenta un conjunto importante de datos sobre su validez externa, cuyas principales fuentes de datos son: (1) las correlaciones entre las calificaciónes compuestas, las calificaciones de los subtests de la DAS y las puntuaciones obtenidas en otras baterías de habilidades múltiples como las escalas de Stanford-Binet y de Wechsler; (2) las correlaciones entre las calificaciones de los subtests de la DAS con las calificaciones escolares y con las puntuaciones de otras pruebas de habilidades cognoscitivas específicas y de aprovechamiento académico, como el Test de Vocabulario con Ilustraciones de Peabody, Revisado (*Peabody Picture Vocabulary Test-Revised*) o PPVT-R y los Tests de Habilidad de Lectura de Woodcock, Revisados (*Woodcock Reading Mastery Tests-Revised*) o WRMT-R, y (3) investigaciones de los perfiles de calificación de la DAS en poblaciones especiales, como individuos sobredotados, con problemas de aprendizaje o mentalmente retardados. Todas esas fuentes de datos generalmente apoyan la estructura jerárquica de la DAS y la posibilidad de comparar las calificaciones compuestas de los subtests y de las pruebas de aprovechamiento con otras medidas similares. La validez de los subtests de diagnóstico para identificar a subgrupos de niños con problemas de aprendizaje parece prometedora, pero se requiere de mayor investigación.

Evaluación general. Como otros han observado (Aylward, 1992; Reinehr, 1992), la complejidad de los procedimientos de aplicación y calificación de la DAS puede obstaculizar su difusión y uso en los ambientes aplicados. Más aún, la DAS es un instrumento relativamente nuevo que no ha sido probado clínicamente, por lo que tendrá que demostrar su valor en la práctica. Otra limitación tiene que ver con los individuos que corresponden al menor y mayor grupos de edad cubiertos por la DAS (es decir, dos años con seis meses y 17 años con 11 meses), para los cuales pueden resultar insuficientes los niveles de piso y de techo de la dificultad de los reactivos. No obstante, la DAS es, en su tipo, un instrumento de vanguardia y, como tal, no superado en cuanto a las posibilidades y ventajas que brinda a sus usuarios. Su estructura jerárquica,

la diversidad de habilidades que cubre y la confiabilidad de sus evaluaciones ofrecen a los usuarios una flexibilidad sin precedentes. En particular, los examinadores pueden elegir de entre una amplia gama de tareas las que resulten más apropiadas para los fines de la evaluación y las necesidades de las personas examinadas. Otra característica destacada de la DAS es la excelente calidad de su *Administration and Scoring Manual* (Manual de aplicación y calificación) y de su *Introductory and Technical Handbook* (Manual introductorio y técnico) (Elliott, 1990a, 1990b). En particular, el primero es de gran amplitud y presenta mucha información con una precisión, claridad y concisión poco comunes, por lo que resulta útil sobre todo para los usuarios potenciales, en especial para los que desean familiarizarse con los avances teóricos y prácticos en el campo de la inteligencia y su medición, de los que la DAS es excelente descripción y representante.

SISTEMA DE EVALUACIÓN COGNOSCITIVA DE DAS-NAGLIERI

El Sistema de Evaluación Cognoscitiva (*Cognitive Assesment System*, CAS) de Das-Naglieri, publicado a finales de los noventa, es otro instrumento reciente de gran importancia para la evaluación individual del funcionamiento cognoscitivo. Esta prueba, desarrollada a lo largo de más de una década, se basa en el modelo PASS de inteligencia propuesto por sus autores (J. P. Das, Naglieri y Kirby, 1994; Naglieri y Das, 1990, 1997a, 1997b). A su vez, el modelo se fundamenta en la teoría de la cognición y organización cerebral expuesta por el neuropsicólogo ruso A. R. Luria.

Las tareas del CAS se diseñaron para medir las funciones cognoscitivas básicas que participan en el aprendizaje, pero se suponen independientes de la escolaridad. Incluyen Planeación, Atención y Procesamiento Simultáneo y Sucesivo. El sistema utiliza pruebas verbales y no verbales mediante los canales sensoriales auditivo y visual. Las de Planeación requieren la evaluación de las estrategias empleadas por el examinado al realizar las tareas. El CAS puede aplicarse a individuos de cinco a 17 años con 11 meses y se diseñó específicamente para vincular la evaluación con la intervención. Debido a sus ventajas teóricas y empíricas y a la cuidadosa estandarización a gran escala a la que se ha sometido, la aparición del CAS ha sido muy esperada por muchos usuarios. De hecho, su versión de prueba ya ha pasado por varias revisiones publicadas (Lambert, 1990; Telzrow, 1990). De acuerdo con la evidencia preliminar disponible sobre la validez del CAS, parece que esta prueba se convertirá en una herramienta importante e innovadora para la evaluación del estado cognoscitivo.

Pruebas para poblaciones especiales

En este capítulo revisamos principalmente escalas individuales y colectivas desarrolladas como pruebas de ejecución, no lingüísticas o no verbales para personas a las que no se puede examinar adecuadamente con los instrumentos individuales que estudiamos en el capítulo anterior o con las pruebas colectivas que veremos en el próximo.

En conjunto, las *pruebas de ejecución* se caracterizan por la manipulación de objetos y un uso mínimo de lápiz y papel; por su parte, las pruebas *sin lenguaje* no requieren que examinador o examinado empleen lenguaje oral o escrito e incluso es posible dar las instrucciones mediante demostraciones, gestos o pantomima. El prototipo de las pruebas colectivas sin lenguaje es la Prueba Beta del Ejército, un instrumento que durante la Primera Guerra Mundial se aplicó a reclutas iletrados o a los que no hablaban inglés (Yerkes, 1921) y del que luego se hicieron revisiones para su uso civil. Sin embargo, al aplicar la prueba no es necesario eliminar del todo el lenguaje, puesto que los examinados suelen tener cierto conocimiento de una lengua común; además, en general es posible traducir algunas instrucciones sencillas y breves o bien presentarlas sucesivamente en dos idiomas sin que ello altere de manera apreciable la naturaleza o dificultad de la prueba. Con todo, ninguno de estos instrumentos requiere que los propios examinados utilicen el lenguaje escrito o hablado.

Otra categoría afín es la de las *pruebas no verbales* o pruebas sin lectura, entre las que se encuentra la mayor parte de las pruebas para escolares, preescolares y para analfabetas de cualquier edad. Si bien estos instrumentos no requieren que el examinado lea o escriba, sí implican que el examinador proporcione instrucciones y comunica-

ciones orales. Además, a menudo miden la comprensión verbal (como el reconocimiento del vocabulario y la comprensión de enunciados y párrafos breves) mediante el uso de reactivos pictóricos que se complementan y acompañan de instrucciones orales. A diferencia de las pruebas no lingüísticas, éstas pueden ser inadecuadas para los individuos que tienen deficiencias auditivas o que no hablan el idioma en el que están escritas.

Si bien las distinciones tradicionales entre las pruebas de ejecución, no lingüísticas y no verbales facilitan la comprensión de sus propósitos, éstas se han desvanecido al aumentar el número de baterías de pruebas que pertenecen a las tres categorías. La combinación de pruebas verbales y de ejecución de las escalas de Wechsler es un ejemplo clásico de dichas baterías.

En este capítulo a las pruebas no las clasificamos en términos de su contenido o procedimientos de aplicación, sino en función de sus usos principales, según lo cual podemos reconocer cuatro categorías principales: pruebas para infantes y de nivel preescolar, para personas con retardo mental, para individuos con discapacidades sensoriales y motrices y las pruebas diseñadas para aplicarse entre culturas o subculturas. No obstante, dicho esquema de clasificación debe permanecer flexible, ya que algunas de las pruebas han demostrado ser de utilidad en más de un contexto. Tal es el caso de algunos instrumentos que originalmente se diseñaron para la examinación transcultural y que ahora son de uso común en los exámenes clínicos.[1]

Por último, aunque algunas de las pruebas que revisaremos en este capítulo se diseñaron como pruebas colectivas, es frecuente que se apliquen individualmente, como sucede en la evaluación clínica para complementar el tipo común de tests de inteligencia y proporcionar una imagen más completa del funcionamiento intelectual del individuo.

Otras permiten la clase de observaciones cualitativas que se asocian con los exámenes individuales por lo que exigen una considerable sofisticación clínica para realizar una interpretación detallada de la prueba de ejecución. En conjunto, están más próximas a las pruebas individuales que revisamos en el capítulo 8 que a las colectivas que veremos en el capítulo 10.

PRUEBAS DE INFANTES Y PREESCOLARES

Todas las pruebas diseñadas para infantes y preescolares deben aplicarse individualmente. Si bien en el jardín de niños es posible examinar a los infantes en pequeños grupos con algunos instrumentos elaborados para el nivel de primaria, en general las pruebas colectivas sólo pueden aplicarse cuando los niños llegan a la edad escolar. Casi todas las pruebas para menores de seis años son verbales o de ejecución, y algunas requieren la manipulación rudimentaria de lápiz y papel.

[1] Véase Sattler (1988, capítulos 12, 14 y 15) para mayor información, evaluación y referencias pertinentes a muchas de las pruebas ilustradas en este capítulo.

Se acostumbra subdividir los primeros cinco años de vida en los periodos infantil y preescolar. El primero se extiende desde el nacimiento a los 18 meses aproximadamente y el segundo de los 18 a los 60 meses. Desde el punto de vista de la aplicación de la prueba, el examen debe hacerse mientras el infante yace de espaldas o mientras una persona lo sostiene sobre su regazo o de alguna otra manera, como veremos luego en el capítulo. Aunque el desarrollo del lenguaje del niño proporciona datos importantes, se usa muy poco el habla para dar instrucciones. Muchas de las pruebas abordan el desarrollo sensoriomotor evidenciado en la capacidad del niño para sostener la cabeza, girar, alcanzar y tomar cosas, así como seguir con los ojos un objeto en movimiento. Por su parte, el preescolar puede caminar, sentarse a la mesa, usar sus manos para manipular los objetos de prueba y comunicarse mediante el lenguaje; mientras que para el infante el examinador constituye el medio de obtener objetos de estímulo, para el preescolar es un individuo, lo que hace de la examinación un proceso más personal en el que por ende aumentan tanto las oportunidades como las dificultades.

Para que el examen psicológico de los niños pequeños sea apropiado es necesario cubrir una amplia gama de conductas que incluye los rasgos emocional y social, así como las habilidades motoras, lingüísticas y cognoscitivas. Además, cada vez se reconoce más la necesidad de considerar la naturaleza del entorno en la evaluación de los niños (Vázquez Nuttall, Romero y Kalesnik, 1992), orientación ecológica que se refleja en algunos de los instrumentos que estudiaremos en el capítulo. En esta sección revisaremos las escalas más comunes diseñadas para la niñez y la niñez temprana, que representan diversas aproximaciones. La Escala de Inteligencia para Edades Preescolar y Primaria de Wechsler, Revisada, pertenece a esta categoría, pero la incluimos en el capítulo 8 para mantener la continuidad de la serie de Wechsler; lo mismo sucedió con las escalas de Stanford-Binet, la Batería de Evaluación para Niños de Kaufman y las Escalas de Habilidad Diferencial, instrumentos que además de emplearse en la evaluación de escolares se utilizan en la de niños preescolares porque abarcan también el periodo de los dos a los seis años.

Antecedentes históricos de la evaluación de infantes y preescolares. La serie de estudios longitudinales que Arnold Gesell y sus colegas realizaron en Yale (Ames, 1989) constituyen uno de los primeros intentos sistemáticos por comprender el desarrollo de infantes y preescolares normales. Estos estudios, que se extendieron a lo largo de cuatro décadas, fructificaron en los Programas de Desarrollo de Gesell cuya primera publicación (Gesell *et al.*, 1940) representó un esfuerzo pionero por proporcionar un método sistemático y empírico de evaluar el desarrollo conductual de los niños pequeños. La mayor parte de los datos de los programas se obtuvo mediante la observación directa de las respuestas del niño a juguetes y otros objetos de estímulo y se complementó con la información proporcionada por los padres o la persona encargada de su cuidado. A lo largo del tiempo, psicólogos y pediatras han hecho uso de los programas de Gesell lo mismo en programas de investigación que en la práctica profesional y, dado que han sido revisados y actualizados por otros investigadores, todavía se les utiliza para complementar los exámenes médicos, especialmente para la identificación temprana de defectos neurológicos y anormalidades conductuales con causa

orgánica.[2] Aunque en la mayor parte de los medios clínicos los programas han sido sustituidos por instrumentos psicométricos más recientes y elaborados, los reactivos y procedimientos que Gesell y sus colegas promovieron han sido incorporados en la mayor parte de las escalas de desarrollo diseñadas para el nivel infantil.

Gracias a la rápida difusión de los programas educativos para niños con retardo mental y al desarrollo de programas preescolares de educación compensatoria para niños con desventajas culturales las décadas de los sesenta a los noventa atestiguaron el renacimiento del interés en las pruebas para infantes y preescolares. Más recientemente, en los Estados Unidos se han promulgado leyes encaminadas a la identificación y remedio precoz de todo tipo de discapacidades físicas y mentales en infantes y preescolares. Algunas de esas leyes (por ejemplo, P. L. 99–457) constituyen enmiendas o extensiones de Ley de Educación para Todos los Niños Discapacitados (*Education for All Handicapped Children Act*, P. L. 94–142), que revisaremos con más detalle en este capítulo. Con el objeto de satisfacer esas apremiantes necesidades prácticas han aparecido nuevas pruebas y publicaciones y se ha llevado a cabo una investigación considerable sobre las aproximaciones innovadoras a la evaluación.[3]

Pruebas estandarizadas del desarrollo infantil temprano

Escalas de Bayley sobre el desarrollo infantil. En la figura 9.1 se ilustran las Escalas de Bayley sobre el Desarrollo Infantil, una prueba especialmente bien elaborada para las edades más tempranas y de la que ahora se cuenta con la segunda edición (Bayley-II, Bayley, 1993). Estas escalas, publicadas originalmente en 1969, incorporan algunos reactivos de los programas de Gesell y de otras pruebas para infantes y preescolares y son, además, el producto final de una investigación realizada durante muchos años por Bayley y sus colegas en la que se incluyen las investigaciones longitudinales del Estudio de Berkeley sobre el Desarrollo (*Berkeley Growth Study*).

La escala Bayley-II proporciona tres herramientas complementarias para evaluar el estado de desarrollo de niños entre el mes y los tres años y medio de edad: la *Escala Mental*, la *Escala Motora* y la *Escala de Clasificación de Conducta*. La Escala Mental prueba funciones como la agudeza sensorial y perceptual, la memoria, el aprendizaje, la resolución de problemas, la vocalización, el inicio de la comunicación verbal y el pensamiento abstracto rudimentario. La Escala Motora ofrece medidas de habilidades motrices gruesas, como sentarse, erguirse, caminar y subir escaleras, así como la destreza de manos y dedos para la manipulación; asimismo, incluye reactivos que evalúan la integración sensorial y perceptual-motora. En el nivel infantil, el desarrollo loco-

[2] El manual más reciente de la versión revisada de los programas originales de Gesell se encuentra en Knobloch, Stevens y Malone (1980). Aunque otras pruebas llevan en su título el nombre de Gesell, ninguna cubre el periodo de la infancia (en Test in Print IV, TIP-IV puede encontrar una lista de las versiones actuales y en la novena edición de los Anuarios de Medición Mental, conocidos como MMY (por *Mental Measurement Yearbook*), aparecen las revisiones de algunas de ellas).

[3] Puede encontrar una historia breve pero informativa de la evaluación preescolar en M. F. Kelley y Surbeck (1991). Para otros materiales relevantes sobre la examinación de infantes y preescolares, véase Aylward (1994), Bracken (1991b), Culbertson y Willis (1993), Kamphaus (1993), C. R. Reynolds y Kamphaus (1990a) y Vázquez Nutall, Romero y Kalesnik (1992).

Figura 9.1. Examen del Infante como se ilustra en la segunda edición del Manual de la Escala de Bayley sobre el Desarrollo Infantil.

(Adaptado de Bayley, 1993, pp. 143, 145 y 150. Copyright © 1993 de The Psychological Corporation. Reproducido con autorización.)

motor y manipulatorio cumple una función importante en las interacciones del niño con su ambiente y por ende en el desarrollo de sus procesos mentales. La Escala de Clasificación de Conducta está diseñada para evaluar diversos aspectos del desarrollo de la personalidad, como la conducta emocional y social, la activación, el tiempo que puede mantenerse la atención, la persistencia y la dirección hacia una meta. La escala ofrece para cada reactivo un sistema de calificación de cinco puntos y descriptores específicos de la conducta valorada. Después de aplicar las otras dos escalas, el examina-

dor la completa de acuerdo con sus propias impresiones y la información proporcionada por la persona encargada del cuidado del niño.

Las escalas de Bayley destacan entre las pruebas para el nivel infantil en lo que respecta a la calidad técnica de los procedimientos con que se elaboraron. Las normas se establecieron en 1 700 niños divididos en 17 grupos de edad entre uno y 42 meses, cada uno constituido por 50 mujeres y 50 varones. La muestra de estandarización se eligió de modo que fuera representativa de la población estadounidense en términos de etnicidad, las principales regiones geográficas y la educación de los padres. Sólo se consideró para la muestra a los niños normales, definidos como los nacidos entre las 36 y las 42 semanas luego de la gestación, que no sufrían de complicaciones médicas importantes y que no tenían una historia de tratamiento para problemas mentales, físicos o conductuales. Las escalas Mental y Motora producen índices separados de desarrollo que se expresan como calificaciones estándares normalizadas con una media de 100 y una *DE* de 15. Esos índices se encuentran en el propio grupo de edad del niño, que hasta los 36 meses se clasifica a intervalos de un mes y posteriormente de tres meses. La Escala de Clasificación Conductual proporciona calificaciones en rangos percentilares que, a su vez, se clasifican como "no óptimo", "cuestionable" o "dentro de los límites normales". En una revisión reciente de varias escalas para la evaluación de niños preescolares, la Bayley-II resultó ser una de dos pruebas que, para la mayor parte de los criterios, cumple las normas de adecuación técnica[4] (D. P. Flanagan y Alfonso, 1995).

Bayley señaló que, al igual que todas las pruebas para infantes, sus escalas no deben usarse para predecir los niveles subsecuentes de habilidad, sino para evaluar el estado actual de desarrollo. Como veremos en el capítulo 12,[5] a tan temprana edad el desarrollo de las habilidades es susceptible de la intervención de muchas influencias, lo que hace que las predicciones a largo plazo resulten de poco valor.

Desde la publicación de las escalas originales de Bayley, el punto central de la evaluación infantil ha cambiado de la evaluación de infantes normales a la de niños en riesgo de sufrir retardos en el desarrollo o que ya los presentan. Aunque se requiere de mayor investigación sobre la utilidad clínica de la Bayley-II, las escalas ya incluyen reactivos que establecen una discriminación máxima entre muestras normales y clínicas y ofrecen algunos datos sobre poblaciones clínicas especiales, característica que les permitirá demostrar su utilidad en la detección temprana de defectos sensoriales y neurológicos, trastornos emocionales y deficiencias ambientales. Además, Aylward (1995) elaboró el Detector del Desarrollo Neurológico Infantil de Bayley (*Bayley Infant Neurodevelopmental Screener*, BINS), una prueba diseñada para hacer una evaluación rápida a infantes de 3 a 24 meses que usa combinaciones de 11 a 13 reactivos del Bayley-II y otras pruebas neurológicas.

Escalas de habilidades infantiles de McCarthy. A nivel preescolar, las Escalas de Habilidades Infantiles de McCarthy (*EHIM, McCarthy Scales of Children's Abilities*,

[4] "La otra fue la Batería Psicoeducativa Revisada de Woodcock-Johnson: Tests de Habilidad Cognoscitiva (*Psycho-Educational Battery-Revised: Tests of Cognitive Ability*, Woodcock y Johnson, 1989, 1990).

[5] Para una revisión de los usos y las limitaciones de los tests de inteligencia infantil, véase Goodman (1990). En Rovee-Collier y Lipsitt (1992) se encuentra una serie de artículos sobre las propiedades psicométricas de las escalas originales de Bayley.

MSCA; McCarthy, 1972) son un buen instrumento para niños de dos y medio a ocho años y medio. El examinador tiene muchas oportunidades de observar la forma en que el niño se acerca a diversos problemas y estímulos, ya que el instrumento está constituido por 18 pruebas agrupadas en seis escalas que se traslapan: Verbal, de Ejecución Perceptual, Cuantitativa, Cognoscitiva General, de Memoria y Motora. El Índice Cognoscitivo General (ICG) se basa en 15 de las 18 pruebas de la batería por lo que se asemeja a las tradicionales medidas globales del desarrollo intelectual. Es una calificación estándar normalizada que se presenta en las mismas unidades que el CI tradicional (con una media de 100 y una *DE* de 16) y se encuentra dentro de cada grupo de tres meses de edad. En el desarrollo de la EHIM se evitó deliberadamente utilizar el término CI por sus connotaciones engañosas, y por lo tanto se describe al ICG como indicador del funcionamiento del niño en el momento del examen, sin implicar inmutabilidad o etiología. Las calificaciones de las cinco escalas adicionales se basan en los mismos grupos de edad y tienen una media de 50 y una *DE* de 10.

Desde la publicación de este instrumento, diseñado para adecuarse a la evaluación cognoscitiva de niños pequeños, se ha acumulado una vasta investigación realizada durante más de dos décadas con las escalas de McCarthy. De particular utilidad resultan los estudios realizados con niños de minorías étnicas y detallados por Valencia (1990), así como la gran cantidad de datos que él y Bracken (1991a) reunieron sobre la validez del instrumento. En lo que respecta al uso clínico de las escalas de McCarthy, la guía preparada por Kaufman y Kaufman (1977) continúa siendo una ayuda indispensable. La visión que prevalece entre muchos revisores es que, a pesar de algunos defectos, las escalas son eficaces y útiles. Las propiedades psicométricas de las escalas son adecuadas para casi todos, especialmente para quienes se encuentran en el rango intermedio de sus poblaciones objetivo.

Escalas piagetianas

Aunque pueden aplicarse mucho después del nivel preescolar, las escalas modeladas en función de las teorías de Jean Piaget sobre el desarrollo han mostrado su mayor utilidad en la niñez temprana. Todas ellas se encuentran en una etapa experimental, por lo que son pocas las que están disponibles comercialmente. Casi todas se elaboraron para utilizarse en los programas de investigación de sus autores, aunque algunas de esas escalas están al alcance de otros investigadores. La principal contribución de las escalas piagetianas al examen psicológico de niños consiste en que pueden aportar un marco teórico que da cuenta de las secuencias del desarrollo de los procesos de pensamiento y un procedimiento de evaluación que se caracteriza por su flexibilidad e interpretación cualitativa.

En el capítulo 3 vimos algunas características de las escalas piagetianas, y en especial nos referimos a la interpretación normativa del desempeño. Básicamente, las escalas piagetianas son ordinales en el sentido de que presuponen una secuencia uniforme del desarrollo por etapas sucesivas que se extienden desde la infancia hasta después de la adolescencia y que se conocen como las etapas sensoriomotora, preoperacional, de las operaciones concretas y de las operaciones formales. Las escalas se ubican además en la aproximación "referida a criterio" en la medida en que proporcionan descripciones

cualitativas de lo que el niño realmente puede hacer. Las tareas piagetianas no se centran en la medición de rasgos, sino en el desarrollo a largo plazo de conceptos específicos o esquemas cognoscitivos.[6] En lo que respecta a la aplicación, las escalas piagetianas pretenden hacer que el niño explique un acontecimiento observado y que exprese las razones que fundamentan su explicación. Por lo general, la calificación no se basa en el número o la dificultad de los reactivos resueltos con éxito, sino en la calidad de las respuestas del niño a un número relativamente pequeño de situaciones problemáticas que se le presentan. En relación con esto, resultan de principal interés las creencias erróneas que el niño revela en sus respuestas incorrectas, por lo que el examinador se concentra más en el proceso de resolución de problemas que en el producto.

La examinación piagetiana es muy adecuada para el trabajo clínico, ya que sus procedimientos son muy individualizados. También logró atraer la atención de los educadores porque permite la integración de examinación y enseñanza, pero su uso más frecuente aún se encuentra en la investigación de la psicología del desarrollo. Las pruebas pueden clasificarse en dos categorías: (1) escalas ordinales para la infancia, y (2) tareas diseñadas para evaluar si el niño se encuentra en la etapa preoperacional, de las operaciones concretas o de las operaciones formales. Existen varios ejemplos de cada una y recientemente D. Sexton, Kelley y Surbeck (1990) revisaron su utilización en varios contextos de investigación. Más adelante describiremos una prueba de cada tipo, seleccionada en parte por su disponibilidad.

Las Escalas Ordinales del Desarrollo Psicológico fueron preparadas por Užgiris y Hunt (1975), también se conocen como Escalas de Desarrollo Psicológico Infantil. Están diseñadas para evaluar la adquisición de las capacidades cognoscitivas entre las dos semanas y los dos años de edad, intervalo que abarca aproximadamente lo que Piaget caracterizó como el periodo sensoriomotor y dentro del cual reconoció seis niveles. Para aumentar la sensibilidad de sus instrumentos, Užgiris y Hunt clasificaron las respuestas en más de seis niveles, cuyo número varía de siete a 14 en las diferentes escalas. La serie incluye seis escalas designadas de la siguiente manera:

1. *Permanencia del objeto*. Surge la noción de la existencia independiente de los objetos, la cual se hace evidente cuando el niño sigue visualmente un objeto y lo busca cuando escapa del alcance de su vista al ser ocultado.
2. *Desarrollo de los medios* para alcanzar los fines deseados del desarrollo. El niño utiliza sus manos y otros objetos, como cuerdas, varas, soportes, etc., para alcanzar objetos.
3. *Imitación*. Comprende la imitación vocal y de gestos.
4. *Causalidad operacional*. El niño reconoce y adapta la causalidad objetiva, lo que muestra en respuestas como la observación de sus manos para producir la conducta deseada en una persona o para activar un juguete mecánico.
5. *Relaciones de objetos en el espacio*. El niño coordina los esquemas de mirar y escuchar para localizar objetos en el espacio y comprende relaciones como recipiente, equilibrio, gravedad.

[6] "Esquema", término que se encuentra a menudo en los escritos de Piaget y que significa esencialmente el marco en el que el individuo ajusta los datos sensoriales de entrada.

6. *Desarrollo de los esquemas* para relacionar los objetos. Para responder a los objetos, el niño mira, siente, manipula, deja caer, arroja, etc. y emplea esquemas socialmente instigados que son apropiados para objetos particulares (digamos, "conducir" un carro de juguete, construir con cubos, usar bolitas perforadas, nombrar objetos).

No se proporcionan normas, pero los autores obtuvieron datos sobre las propiedades psicométricas de sus escalas al aplicarlas a 84 infantes, hijos de estudiantes graduados y de miembros del equipo de la Universidad de Illinois. Se señala el grado de acuerdo entre observadores y el acuerdo test-retest después de un intervalo de 48 horas. En general, las pruebas parecen bastante satisfactorias en ambos aspectos. También se informa que los índices de ordinalidad, calculados para cada escala a partir de las calificaciones de los mismos 84 niños, son muy satisfactorios.[7]

Aunque se pretendía que las Escalas Ordinales desarrolladas por Užgiris y Hunt fueran provisionales, se les ha utilizado mucho para fines de investigación.[8] Originalmente se diseñaron para medir los efectos de condiciones ambientales específicas en la tasa y el curso del desarrollo infantil. A este respecto, los estudios realizados con infantes criados en condiciones diferentes y con los participantes en programas de intervención han indicado que esas variables ambientales tienen efectos importantes sobre la edad promedio en que los niños alcanzan los diferentes pasos en las escalas de desarrollo. Los autores (Užgiris y Hunt, 1987) publicaron un texto en el que revisan los estudios acerca del uso de las escalas ordinales en la descripción del desarrollo cognoscitivo de infantes normales y excepcionales. La secuencia de adquisición indicada por las escalas atañe principalmente a las interacciones del infante con objetos inanimados, que a su vez se consideran precursoras del desarrollo de la conducta comunicativa y de otras habilidades adaptativas (Dunst y Gallagher, 1983; Kahn, 1987).

El segundo ejemplo de un instrumento piagetano es el Equipo para la Evaluación del Concepto de Conservación (*Concept Assessment Kit-Conservation*, CAK), una prueba distribuida por una editorial sobre la misma base que otras pruebas psicológicas. Diseñada para niños de cuatro a siete años de edad, evalúa uno de los conceptos piagetanos mejor conocidos, la conservación, que se refiere a la comprensión del niño de que las propiedades de los objetos como su peso, volumen y número permanecen inalteradas cuando sufre transformaciones de forma, posición u otros atributos. Los autores (Goldschmid y Bentler, 1968b) consideran que la conservación indica la transición del niño de la etapa preoperacional a la etapa de operaciones concretas, que Piaget ubica aproximadamente entre los siete y ocho años.

Durante la prueba, el procedimiento es esencialmente el mismo: se muestran al niño dos objetos idénticos; luego, el examinador hace ciertas transformaciones en uno de

[7] Los procedimientos para la medición de la ordinalidad y la aplicación del análisis de escalogramas a las escalas piagetanas son polémicos, hecho que debe recordarse al interpretar e informar de los índices de ordinalidad (véase F. H. Hooper, 1973; A. C. Rosenthal, 1985).

[8] El valor potencial de las escalas en la evaluación clínica también es muy reconocido; la publicación de un manual y de formas de calificación especialmente diseñadas para el uso clínico y educativo de las escalas constituyeron un paso significativo hacia el reconocimiento de ese potencial (Dunst, 1980).

ellos y lo interroga sobre su semejanza o diferencia, y después de que contesta le pide que explique su respuesta. En cada reactivo se otorga un punto por el juicio correcto y otro por una explicación aceptable; por ejemplo, el examinador empieza con dos vasos comunes que contienen igual cantidad de agua (cantidad continua) o de granos de maíz (cantidad discontinua) y vierte el contenido en un plato plano o en varios vasos pequeños. En otra tarea, el examinador muestra al niño dos bolas iguales de plastilina y luego aplana una de ellas en una torta y pregunta si la bola pesa tanto como la torta.

Se dispone de tres formas de la prueba, las formas A y B son paralelas y cada una presenta seis tareas de conservación: Espacio Bidimensional, Número, Sustancia, Cantidad Continua, Cantidad Discontinua y Peso. La correlación entre las calificaciones de ambas formas es .95. La forma C incluye dos tareas diferentes, Área y Longitud, y su correlación con las formas A y B es de .76 y .74 respectivamente.

Las normas se establecieron en una muestra de estandarización compuesta por 560 niños y niñas entre cuatro y ocho años que asistían a escuelas, guarderías y centros *Head Start* en el área de Los Ángeles, California. Esas normas deben considerarse tentativas en vista del pequeño número de casos en cada grupo de edad y de las limitaciones en la representatividad de la muestra. Tal como lo anticipa la teoría de Piaget, al aumentar la edad se observa un incremento sistemático en las calificaciones promedio de cada grupo, así como un aumento repentino entre los seis y los ocho años.

Los autores hicieron varios análisis estadísticos para evaluar: la confiabilidad (en las puntuaciones, en el retest, en la prueba de formas paralelas y en la de Kuder-Richardson); la ordinalidad o posibilidad de formar escalas, y la composición factorial del instrumento (véase también Goldschmid y Bentler, 1968a). Aunque se basan en muestras pequeñas, los resultados indican una confiabilidad generalmente satisfactoria y una buena evidencia de ordinalidad, así como la presencia de un factor común de conservación en todas las tareas.

Los estudios comparativos realizados en siete países indican que la prueba puede aplicarse en culturas muy diversas con una confiabilidad elevada y tendencias de edad aproximadamente similares (Goldschmid *et al.*, 1973). Pese a ello, se encontraron diferencias entre culturas y subculturas en las edades promedio a las que se adquieren los conceptos, es decir, que las curvas de edad pueden ser desplazadas horizontalmente por uno o dos años (véase también Figurelli y Keller, 1972; Wasik y Wasik, 1971); además, se observó que el entrenamiento en las tareas de conservación mejora significativamente las calificaciones (véase también Goldschmid, 1968; B. J. Zimmerman y Rosenthal, 1974a, 1974b). El manual cita datos interesantes sobre la validez de constructo de la prueba, que también se encuentran en un estudio reciente de F. A. Campbell y Ramey (1990).

Evaluación de la aproximación piagetiana. Aún persiste la controversia sobre las bases teóricas y empíricas de la aproximación piagetiana al desarrollo cognoscitivo (véase, por ejemplo, Inhelder, de Caprona y Cornu-Wells, 1987; Liben, 1983; Sugarman, 1987). Hay interrogantes sobre las implicaciones de los efectos del entrenamiento y sobre el posible efecto de las diferencias transculturales en la interpretación de las etapas piagetianas de desarrollo. Además, al identificar las etapas mediante las escalas ordinales se enfrenta un obstáculo al que los investigadores piagetanos llaman

"*décalage*",[9] o inconsistencia en las secuencias anticipadas, y se dispone de un cuerpo creciente de datos que cuestiona la continuidad y regularidad implicadas del desarrollo intelectual. Con mucha frecuencia la etapa correspondiente a la ejecución de un individuo varía con la tarea, no sólo cuando su solución requiere la participación de procesos diferentes, sino también cuando se aplica el mismo proceso a distintos contenidos (Dasen, 1977; Goodnow, 1976; Horn, 1976; J. McV. Hunt, 1976).

También debe mencionarse que se ha descubierto que las escalas piagetianas tienen una elevada correlación con los tests estandarizados de inteligencia (Gottfried y Brody, 1975; Kaufman, 1971; M. E. Sexton, 1987) y que su correlación con las pruebas de aprovechamiento escolar de los niños de primer grado es similar a la de los tests colectivos de inteligencia (Kaufman y Kaufman, 1972). Este traslape ha sido corroborado por investigadores independientes con distintos instrumentos (Humphreys, Rich y Davey, 1985). Todos estos descubrimientos indican que las escalas piagetianas, los tests estandarizados de inteligencia y las medidas del aprovechamiento escolar tienen mucho en común a pesar de las considerables diferencias en la metodología, además de que cada una aporta elementos únicos y valiosos a la evaluación global de los niños. La aplicación de las escalas piagetianas resulta más difícil y lleva más tiempo, pero aporta una imagen más completa de lo que el niño puede realizar y la forma en que lo hace, especialmente cuando se integran con normas y medidas referidas a dominio (D. Sexton, *et al.*, 1990).

La investigación actual sobre las actividades mentales de los niños pequeños avanza a un ritmo acelerado y los descubrimientos empíricos revisan y amplían los primeros conceptos piagetanos (véase por ejemplo, Butterworth, Harris, Leslie y Wellman, 1991; Whiten, 1991). De hecho, ahora existe una serie de aproximaciones innovadoras, llamadas colectivamente "neopiagetianas", que abordan temas concernientes al desarrollo cognoscitivo desde una perspectiva en la que confluyen la teoría de Piaget y la del procesamiento de información (Beilin y Pufall, 1992; Demetriou, 1988). En el campo de la evaluación, algunos investigadores neopiagetanos combinan varias aproximaciones dinámicas y emplean de manera formal el aprendizaje mediado para evaluar la capacidad mental con una mínima utilización de la base de conocimientos previos del individuo (Pascual-Leone e Ijaz, 1991). Estas técnicas, aún experimentales, pretenden aplicarse lo mismo a niños tan pequeños como de dos o tres años que a individuos de diversos antecedentes culturales, sociales y lingüísticos.

Tendencias actuales en la evaluación de la infancia y la niñez temprana

Históricamente, la validez de los tests de inteligencia se ha vinculado ante todo a los criterios de diferenciación por edad y a las correlaciones con el aprovechamiento escolar. En el caso de los infantes, el progreso adecuado se calculaba casi en exclusiva al comparar sus resultados con las normas correspondientes a su edad en la amplia gama de tareas incluidas en escalas de desarrollo como las de Bayley. Sin embargo, los esfuerzos sociales recientes para la identificación y el remedio precoz de las deficiencias requieren que las herramientas diseñadas para evaluar el funcionamiento cognoscitivo en los infantes tengan poder predictivo. Lo anterior ha tenido como consecuencia

[9] Literalmente, un "desfase" del patrón de respuestas teóricamente esperado.

nuevos intentos por generar instrumentos y procedimientos con suficiente valor predictivo como para resultar de uso práctico, a pesar de las dificultades encontradas cuando se pretende seguir los cambios en el desarrollo de la competencia intelectual a diversos niveles de edad.

Una de las aproximaciones recientes más interesantes consiste en la medición de las habilidades para el procesamiento de información, como lo hace el Test de Inteligencia Infantil de Fagan (*Fagan Test of Infant Intelligence;* Fagan, 1992; Fagan y Detterman, 1992). Esta aproximación se funda en el conocido hecho de que los infantes prefieren la novedad, lo que a su vez permite estudiar su capacidad para abstraer y retener información. El Test de Fagan, diseñado para diferenciar a los niños normales de los que sufren deficiencias cognoscitivas, evalúa la atención visual selectiva a la novedad en infantes de tres a 12 meses. Los estímulos son imágenes de rostros, y la "calificación" se basa en el tiempo que dedican a las imágenes nuevas en contraste con las familiares. En la figura 9.2 se muestra la versión portátil de la presentación de este instrumento, el cual predice la ejecución posterior en tests de inteligencia tan bien o mejor que las medidas estándares de inteligencia infantil. A los tres años, la correlación entre las calificaciones del test de Fagan y el CI fluctúa entre .40 y aproximadamente .60. El instrumento aún está en evolución y ha recibido muchas críticas (véase, por ejemplo, Benasich y Béjar, 1992; Goodman, 1990). Comoquiera que sea, es indudable que se necesitan mayores datos sobre su utilidad clínica para predecir las deficiencias cognoscitivas (incluido el retardo mental), ya que la aproximación que representa tiene una base empírica sólida que es congruente con la información que existe sobre la naturaleza de la inteligencia infantil, que estudiaremos en el capítulo 12.

Cada vez es mayor el reconocimiento de que la eficacia de una intervención depende de que la evaluación de los niños pequeños sea tan amplia como precisa y válida, ya que para todo fin práctico resulta inadecuado confiar en una sola calificación global, sea el famoso CI o un índice de desarrollo. Estas calificaciones pueden servir para clasificar a los niños en categorías, pero no informan a los usuarios sobre sus ventajas y desventajas. La evaluación individualizada de infantes, niños e incluso de personas mayores debe incluir métodos y fuentes de información múltiples de las áreas pertinentes, como las habilidades lingüísticas, motoras y sociales. En respuesta a estas necesidades se preparan nuevos sistemas de evaluación que incluyen procedimientos y profesionales de varias disciplinas.

Un ejemplo de dicha tendencia es la Evaluación del Desarrollo de Infantes y Niños Pequeños (*Infant-Toddler Developmental Assessment,* IDA), que se basa en el trabajo realizado por un equipo multidisciplinario de especialistas en la niñez temprana (Provence, Erikson, Vater y Palmeri, 1995a, 1995b, 1995c). En esencia, la IDA es un marco de referencia que orienta la identificación de los niños (del nacimiento a los tres años) que están en riesgo de presentar retardos en el desarrollo. Sus materiales facilitan la recolección, el registro, la interpretación y la síntesis de datos desde la llegada del niño hasta el desarrollo de un plan. En cada fase, el procedimiento incluye la participación de los padres, la revisión médica y la evaluación del desarrollo basada en la observación y en entrevistas con los padres y otras personas encargadas del cuidado del niño. La parte del IDA que más se asemeja a las medidas tradicionales del desarrollo infantil es el *Perfil del Desarrollo del Nacimiento a los Tres Años de Provence* (*Provence Birth to Three Developmental Profile*), instrumento que utiliza procedimientos

Figura 9.2. Versión Portátil de la Presentación del Test de Inteligencia Infantil de Fagan.
(Tomado de Fagan y Detterman, 1992, p. 189. Copyright © 1992 de Ablex Publishing. Reproducido con autorización.)

estandarizados de aplicación y calificación para evaluar ocho áreas y secuencias de desarrollo y cuyos resultados se refieren a la edad en términos objetivos bien documentados del desarrollo más que en calificaciones estándares, rangos percentilares u otras comparaciones grupales. A este respecto, Sally Provence, la pediatra que elaboró el perfil, siguió la tradición establecida por Gesell en sus programas de desarrollo. La eficacia de la IDA, como la de cualquier otro instrumento, obviamente está vinculada a la capacitación y experiencia de los profesionales que lo emplean y debe evaluarse en la investigación y práctica clínica. No obstante, éste y otros sistemas similares se diseñaron para enfrentar la crítica dirigida a la práctica de confiar en exceso en los tests de inteligencia (véase, por ejemplo, Goodman, 1990) y, si se aplican adecuadamente, pueden demostrar ser de gran valor práctico. Debe advertirse que el uso de la IDA no excluye la aplicación de medidas tradicionales del funcionamiento cognoscitivo o de cualquier otro instrumento que pueda ser necesario para tales fines, como estimar la posición relativa del niño entre otros niños de su edad.

Además del cambio hacia una evaluación más amplia e integrada, existen otras dos tendencias que tienen un impacto importante en la evaluación de niños pequeños

y de la mayor parte de las poblaciones especiales analizadas en este capítulo. Una es la influencia sobre el proceso de evaluación de la postura ecológica, la cual requiere la consideración de varios aspectos del entorno del niño.[10] La segunda atañe al mayor énfasis en el vínculo entre evaluación e intervención, reconocido en el trabajo clínico en el que el diagnóstico está inextricablemente ligado al tratamiento. En el contexto de la intervención temprana y la educación, se considera esencial que los maestros cuenten con guías prescriptivas basadas en el perfil único de capacidades y discapacidades del individuo (Bagnato y Neisworth, 1991; Witt, Elliott, Gresham y Kramer, 1988).

EVALUACIÓN EXTENSA DE PERSONAS CON RETARDO MENTAL

La examinación de niños con discapacidades físicas o mentales mostró un notable crecimiento en los Estados Unidos después de la aprobación en 1975 de la Ley de Educación para Todos los Niños Discapacitados (*Education for All Handicapped Children Act; P.L. 94–142*), conocida ahora como Ley de Educación para los Individuos con Discapacidades (*Individuals with Disabilities Education Act, IDEA*), y de otros estatutos relacionados.[11] Para poner en práctica esa legislación se requieren cuatro condiciones básicas: (1) reconocer a todos los niños con discapacidades mediante el uso de instrumentos de identificación preliminar, (2) que un equipo de especialistas evalúe a todos los niños identificados para determinar las necesidades educativas particulares, (3) que la escuela establezca un programa individualizado de educación para cumplir esas necesidades y (4) que en el curso del programa se reevalúe periódicamente a cada niño. En distintas partes de este libro (capítulos 8 y 17, así como ésta y las dos secciones siguientes) revisamos las pruebas adecuadas para usarse en los programas educativos que cumplen las exigencias legales (véase también Jacobson y Mulick, 1996).

La Asociación Estadounidense para [el estudio] del Retardo Mental (*American Association on Mental Retardation*, AAMR, 1992) afirma en su manual de definición y clasificación que "el retardo mental se refiere a limitaciones sustanciales en el funcionamiento actual. Se caracteriza por un funcionamiento intelectual significativamente inferior al promedio, que coexiste con limitaciones relacionadas en dos o más de las siguientes áreas de habilidades adaptativas: de comunicación, de cuidado personal, para vivir en el hogar, sociales, de uso comunitario, de automanejo, de salud y seguridad, de funcionamiento académico, de manejo del tiempo libre y de trabajo" (p. 1). El manual también estipula que la condición se manifiesta antes de los 18 años. Aunque esta definición no es muy diferente de la anterior (Grossman, 1983, p. 11), el sistema asociado de clasificación cambió en forma radical para dar cabida a la idea de que el retardo mental no es un rasgo, sino una discapacidad que surge de la interacción entre las limitaciones del individuo y las exigencias del ambiente. Mientras que el sistema anterior especifica niveles

[10] En una sección posterior revisaremos la evaluación del ambiente.

[11] Las principales son P.L. 99–457 y P.L. 101–476, aprobadas respectivamente en 1986 y 1990. Para un análisis de las implicaciones que la legislación federal y los casos legales tienen en la examinación y evaluación de los niños, véase Ayers, Day y Rotatori (1990), DeMers, Fiorello y Langer (1992), M. P. Kelly y Melton (1993), y Sattler (1988, pp. 767–784).

de retardo que, de acuerdo con la ejecución en los tests de inteligencia, van de moderado a profundo, el actual clasifica la intensidad de los apoyos que el individuo necesita en cuatro dimensiones (funcionamiento intelectual y habilidades adaptativas, aspectos psicológicos o emocionales, aspectos de la etiología y salud física, y aspectos del ambiente). En la tabla 9.1 se presenta la definición y ejemplos de las intensidades de los apoyos.[12]

La definición actual reafirma la idea de que la limitación intelectual es una condición necesaria pero no suficiente del retardo mental, y, para que éste exista, la limitación intelectual debe afectar las habilidades adaptativas o de afrontamiento. El nivel de umbral de la ejecución en el test de inteligencia que se emplea para demarcar el retardo mental es una puntuación entre 70 y 75 en una escala que tenga una media de 100 y una *DE* de 15, o *aproximadamente* dos o más desviaciones estándares por debajo de la media, esto se tiene en cuenta para el error de medición y acentúa el hecho de que no existen líneas divisorias claramente definidas entre los "retardados mentales" y los "normales". La eliminación de los niveles de retardo y el mayor énfasis en las habilidades adaptativas y los aspectos ambientales se concentran en el patrón único de fortalezas y debilidades presentes en el individuo y sobre la posibilidad de crecimiento. Las revisiones de la AAMR

Tabla 9.1

Definición y ejemplos de la intensidad de los apoyos

Intermitente

El apoyo se presenta "según se requiera". Se caracteriza por la naturaleza episódica: la persona no siempre necesita el o los apoyos, o requiere de apoyos a corto plazo durante las transiciones del ciclo vital (por ejemplo, al perder el empleo o durante una crisis médica aguda). La intensidad de los apoyos intermitentes puede ser elevada o baja.

Limitada

La intensidad del apoyo se caracteriza por su constancia, limitada en el tiempo pero no de naturaleza intermitente, puede requerir la participación de menos miembros del equipo y su costo es menor que el de los niveles más intensos de apoyo (por ejemplo, capacitación para un empleo de tiempo parcial o apoyos en la transición de la escuela al periodo adulto).

Extensa

El apoyo se caracteriza por la intervención regular (es decir, diaria), al menos en algunos entornos (como el trabajo o el hogar), y no se limita en el tiempo (por ejemplo, apoyo a largo plazo y apoyo a largo plazo en el hogar).

Penetrante

El apoyo se caracteriza por ser constante y de alta intensidad, porque se proporciona en todos los ambientes y por su naturaleza potencial de sostener la vida. Generalmente incluye la participación de más miembros del equipo y más intrusión que los apoyos extensos o temporalmente limitados.

(Tomado de AAMR, 1992, p. 26. Copyright © 1992 de la American Association on Mental Retardation. Reproducido con autorización.)

[12] El *Manual Estadístico y de Diagnóstico de Trastornos Mentales* (DSM-IV, 1994) de la Sociedad Estadounidense de Psiquiatría todavía especifica cuatro grados de gravedad del retardo mental de acuerdo con los niveles del CI, a saber, leve, moderado, grave y profundo.

han resultado polémicas, pues mientras algunos críticos señalan que las nuevas especificaciones son imprecisas, no pueden evaluarse confiablemente e incrementan la proporción de la población que puede requerir los servicios escolares especiales (Gresham, MacMillan y Siperstein, 1995; MacMillan, Gresham y Siperstein, 1993; Matson, 1995), otros afirman que esas objeciones carecen de bases empíricas (Reiss, 1994). En cualquier caso, es probable que la adopción de estándares diferentes ocasione la modificación de la forma en la que se realizan las evaluaciones y de sus resultados; sin embargo, es probable que sea demasiado pronto para evaluar los efectos de las revisiones.

Además de los tests individuales de inteligencia, como los que estudiamos en el capítulo 8, los programas de evaluación de las personas con retardo generalmente incluyen la medición de la conducta adaptativa en situaciones cotidianas.[13] El prototipo de las escalas de evaluación de la conducta adaptativa es la Escala de Madurez Social de Vineland (*Vineland Social Maturity Scale*) que desarrolló en los treinta el director de la Escuela de Capacitación de Vineland, Edgar Doll (1935–1965). Como resultado de sus observaciones de las diferencias entre los pacientes con retardo mental, Doll diseñó una forma estandarizada de registro para evaluar el nivel de desarrollo del individuo tanto al ocuparse de sus necesidades prácticas como al asumir las responsabilidades de la vida cotidiana. Se dispone de tres versiones de la última revisión de las Escalas de Conducta Adaptativa de Vineland (*Vineland Social Maturity Scale* VABS; P. L., Harrison, 1985; Sparrow, Balla y Cicchetti, 1984a, 1984b) que pueden usarse de manera independiente o combinada. Dos de las escalas son las Ediciones de Entrevista, con las que se obtiene información en una entrevista semiestructurada con uno de los padres u otra persona encargada del cuidado del individuo. Una de esas versiones es la Forma de Reconocimiento que consta de 297 reactivos y es muy parecida a la versión anterior de Vineland. La otra es la Forma Ampliada compuesta por 577 reactivos y que también ofrece una base sistemática para preparar programas individualizados de educación o tratamiento. Ambas versiones pueden aplicarse desde el nacimiento hasta los 18 años y a adultos de bajo funcionamiento. La tercera versión es la Edición para el Salón de Clases, que abarca de los tres a los 12 años de edad y presenta un cuestionario con 244 reactivos que el maestro debe contestar. Las correlaciones entre las calificaciones de la Edición para el Salón de Clases y la Forma de Reconocimiento varía entre .31 y .54, lo que indica que no pueden usarse indistintamente.

Todas las versiones de las escalas de Vineland se concentran en lo que el individuo *hace* habitualmente más que en lo que *puede* hacer. Los reactivos se clasifican en cuatro áreas adaptativas que se muestran en la tabla 9.2 junto con sus subáreas y una breve descripción de la conducta cubierta. Las dos Ediciones de Entrevista incluyen un conjunto opcional de 32 reactivos que abordan la conducta indeseable o mal adaptada que puede interferir con el funcionamiento del individuo. Todas las versiones incluyen formas bien diseñadas para informar a los padres de los resultados.

Las dos Ediciones de Entrevista se estandarizaron en una muestra estadounidense representativa compuesta por 3 000 individuos, que iban del nacimiento a los 18 años con 11 meses, que se estratificó de acuerdo con el Censo Estadounidense de 1980 sobre la base del género, grupo étnico, tamaño de la comunidad, región y nivel educati-

[13] Debe advertirse que el funcionamiento de las principales escalas individuales de inteligencia no es igualmente bueno cuando se aplican a personas con retardo mental (véase, por ejemplo, Spruill, 1991).

Tabla 9.2

Contenido de las Escalas de Conducta Adaptativa de Vineland

Áreas y subáreas	Descripción
Comunicación	
Receptiva	Lo que el individuo entiende
Expresiva	Lo que el individuo dice
Escrita	Lo que el individuo lee y escribe
Habilidades de la vida cotidiana	
Personal	La forma en que el individuo come, se viste y practica la higiene personal
Doméstica	Las tareas domésticas que el individuo realiza
Comunitaria	La forma en que el individuo emplea el tiempo el dinero, el teléfono y las habilidades laborales
Socialización	
Relaciones interpersonales	La forma en que el individuo interactúa con los demás
Juego y tiempo libre	La forma en que el individuo juega y usa su tiempo libre
Habilidades de afrontamiento	La forma en que el individuo demuestra responsabilidad y sensibilidad hacia los demás
Habilidades motoras	
Gruesa	La forma en que el individuo utiliza brazos y piernas para el movimiento y la coordinación
Finas	La forma en que el individuo utiliza manos y dedos para manipular objetos
Conducta adaptativa compuesta	Compuesta por las cuatro áreas mencionadas antes
Conducta inadaptada[a]	Conductas indeseables que pueden interferir con el funcionamiento adaptativo

[a] Se incluye para uso opcional en la Forma de Reconocimiento y en la Forma Ampliada pero no en la Edición para el Salón de Clases.

(Adaptada de Sparrow, Balla y Cicchetti, 1984a, p. 3. Copyright © 1984, 1985 de la American Guidance Service, Inc. Reproducida con autorización del editor. Todos los derechos reservados.)

vo de los padres. También se establecieron normas complementarias en grupos especiales que incluían muestras de adultos internados y no internados con retardo mental y muestras de niños internados con perturbaciones emocionales, deficiencias visuales y deficiencias auditivas. La Edición para el Salón de Clases se estandarizó en una muestra aproximada de 3 000 niños de tres a 12 años con 11 meses que asistían a la escuela en 38 estados de los Estados Unidos. En este caso, la muestra se estratificó sobre las mismas bases que las otras ediciones.

Para las tres formas de las cuatro calificaciones de dominio y para el Compuesto de la Conducta Adaptativa se proporcionan calificaciones estándares con una media de 100 y una *DE* de 15. Se presentan las bandas de error (basadas en el EEM) para cinco diferentes niveles de confianza (de 68 a 99 por ciento). Para las mismas calificaciones re-

sumidas se dispone también de rangos percentilares, estaninas, equivalencias de edad y niveles de adaptación (es decir, categorías cualitativamente descriptivas). Para las calificaciones de las subáreas, los resultados pueden expresarse como niveles de adaptación y como equivalentes de edad. En el caso de la conducta mal adaptada, los resultados se expresan sólo en niveles cualitativos (inadaptados). Las normas complementarias pueden utilizarse para encontrar los rangos percentilares y los niveles de adaptación dentro de cada grupo especial apropiado. Existen programas de cómputo que convierten las calificaciones no procesadas en calificaciones derivadas y realizan el análisis de los perfiles.

Para todas las ediciones, la mediana de los coeficientes de confiabilidad por consistencia interna para las calificaciones de dominio y compuesta se encuentra mayoritariamente por arriba de .90. Como era de esperarse, la confiabilidad es menor en las subáreas y varía considerablemente según el nivel de edad y el área de contenido, aunque la mediana de la confiabilidad de las subáreas se localiza entre .70 y .80. Los datos sobre la confiabilidad en el retest y entre calificadores indica que existe una buena estabilidad en intervalos breves y un acuerdo satisfactorio cuando dos entrevistadores distintos aplican la escala a los mismos individuos.

Diversas clases de datos resumidos en los manuales de las tres formas contribuyen a la validación de constructo. Hasta cierto punto, la validez se inicia en la formulación de los constructos de conducta adaptativa que dirigen la preparación y selección de reactivos. Los datos de validación empírica se proporcionan en los análisis de las muestras de estandarización y en los estudios realizados por investigadores independientes que incluyen datos sobre las tendencias de desarrollo en áreas y subáreas de ejecución, análisis factoriales de las calificaciones de áreas y subáreas, comparaciones de los perfiles de calificación obtenidos por muestras de individuos retardados y discapacitados incluidas en las normas complementarias y correlaciones con otros inventarios de conducta adaptativa y con tests de habilidad como la escala WISC-R, la batería K-ABC y el Test de Vocabulario con Ilustraciones de Peabody (*Peabody Picture Vocabulary Test*, PPVT-R).

En general, en el desarrollo y evaluación de las escalas de Vineland se utilizan procedimientos de elevada calidad técnica que se describen de manera clara y completa en los manuales a la vez que reflejan los avances logrados desde la publicación de la edición anterior. No obstante, la efectividad práctica del instrumento depende de que se tenga un conocimiento pleno de las características psicométricas descritas en el manual y en la bibliografía.[14]

Existe gran interés en el uso de los resultados de evaluación como base para diseñar y elegir los programas apropiados de capacitación para las personas con retardo mental. Este interés lleva, a su vez, al desarrollo de un mayor número de escalas para medir la conducta adaptativa.[15] Un ejemplo son las Escalas de Conducta Adaptativa (*Adaptive Behavior Scales*, ABS) elaboradas por la AAMR con los mismos objetivos generales que las escalas de Vineland. La segunda edición de la Escala de Conducta Adaptativa Institucional y Comunitaria de la AAMR (*AAMR Adaptive Behavior Sca-*

[14] Véase, por ejemplo, Middleton, Keene y Brown (1990), Poth y Barnett (1988), Raggio y Massingale (1990), Schatz y Hamdan-Allen (1995), y Silverstein (1986). Para otra descripción y una evaluación independiente de las tres formas de las escalas de Vineland, véase I.A. Campbell (1985) y C. R. Reynolds (1986).

[15] Para una revisión de varias de esas escalas, véase Fox y Meyer (1990), Knoff (1992) y Sattler (1988, capítulo 15).

le-Residential and Community, Second Edition, ABS-RC:2; Nihira, Leland y Lambert, 1993) se estandarizó con 4 000 adultos con discapacidades en el desarrollo que vivían en medios comunitarios o institucionales. Tiene calificaciones en 18 áreas, de las cuales 10 tienen que ver con las habilidades de afrontamiento y ocho con la conducta social, e incluye diversos tipos de patrones inadaptados. Por otro lado, las normas de la segunda edición de la Escala de Conducta Adaptativa en la Escuela de la AAMR (*AAMR Adaptive Behavior Scale-School, Second Edition, ABS-S:2;* Lambert, Nihira y Leland, 1993) se establecieron en estudiantes con y sin retardo mental de tres a 18 años.

Otra área que debe evaluarse en los individuos con retardo mental es el desarrollo motor (que también se examina en las escalas infantiles). El prototipo de los instrumentos con ese propósito son los Tests de Destreza Motora de Oseretsky (*Oseretsky Tests of Motor Proficiency*) publicados originalmente en Rusia en 1923. Otras aplicaciones de los tests Oseretsky se encuentran en el examen de niños con discapacidades motoras, disfunción cerebral mínima o problemas de aprendizaje, sobre todo en relación con la aplicación de programas individualizados de entrenamiento. La versión actual de las escalas Oseretsky es el Test de Destreza Motora de Bruininks-Oseretsky (*Bruininks-Oseretsky Test of Motor Proficiency*, Bruininks, 1978). La batería, cuya solución requiere de 45 a 60 minutos, comprende 46 reactivos agrupados en ocho subtests y produce tres calificaciones: un Compuesto Motor Grueso que mide la ejecución de los músculos largos de hombros, tronco y piernas; un Compuesto Motor Fino que mide la ejecución de los músculos pequeños de los dedos, manos y antebrazos; y un Compuesto Total de la Batería. También existe una Forma Corta con 14 reactivos que requiere de 15 a 20 minutos y proporciona un solo índice de la destreza motora general. La ejecución puede expresarse en términos de calificaciones estándares basadas en edad, rangos percentilares y estaninas; para cada subtests también se dispone de equivalentes por edad. La batería se estandarizó con una muestra de 765 niños de cuatro y medio a 14 años y medio, elegidos de modo que fueran representativos de la población estadounidense. La confiabilidad test-retest de las calificaciones compuestas, en intervalos de siete a 12 días, se agrupa alrededor de .80. Se investigó la validez de diversas formas, incluyendo un análisis factorial de los reactivos, una diferenciación por edad y comparaciones entre niños con y sin retardo mental y problemas de aprendizaje.

Uno de los retos principales en la evaluación del retardo mental consiste en distinguir entre esta condición y las demoras del desarrollo, especialmente durante la infancia y la niñez temprana. Como hemos visto, la evaluación cognoscitiva en este periodo no sólo es menos confiable que en otras edades, sino que también existe la posibilidad de que lo que parece un funcionamiento cognoscitivo deteriorado sea resultado de otras condiciones (Hodapp, Burack y Zigler, 1990). Entre los factores principales que pueden influir de manera negativa en el nivel de capacidad intelectual y habilidades adaptativas del niño se encuentran las discapacidades sensoriales y motoras y un entorno familiar adverso. El resto del capítulo incluye la revisión de temas relacionados con esos dos factores que, por supuesto, pueden presentarse por separado o en combinación.

EXAMINACIÓN DE PERSONAS CON DISCAPACIDADES FÍSICAS

Aunque los problemas que supone la examinación de personas discapacitadas han recibido atención durante varias décadas, la legislación promulgada en los Estados Uni-

dos durante la década de los setenta estimuló la realización de esfuerzos especiales para resolverlos. La obligación de proporcionar educación adecuada a todos los niños físicamente discapacitados se encuentra señalada en la Ley de Educación para los Individuos con Discapacidades. A un nivel mucho más amplio, los derechos civiles generales reconocidos para otras minorías se extendieron para abarcar a las personas con discapacidades físicas, primero gracias a la sección 504 de Ley de Rehabilitación de 1973 y, más recientemente, por la Ley de Estadounidenses con Discapacidades de 1990 (ADA-P.L. 101–336).[16] Esas leyes prohíben la discriminación en las áreas de: (1) prácticas de empleo, (2) acceso físico a los servicios públicos, (3) educación preescolar, primaria y secundaria, (4) educación posterior a la secundaria y (5) salud, bienestar y servicios sociales. La ADA refuerza mandatos anteriores y los extiende a organizaciones en el sector privado.

El examen de niños con discapacidad física a una edad temprana es especialmente importante para proporcionar desde el inicio las experiencias educativas apropiadas, lo que ayuda a prevenir la acumulación de deficiencias de aprendizaje que podría magnificar los efectos de las discapacidades en el desarrollo intelectual.[17] A cualquier edad, el examen de personas con discapacidades físicas plantea problemas especiales en lo que respecta a la aplicación de la prueba y a la interpretación adecuada de sus resultados. Hasta ahora, las principales formas de manejar dicha examinación incluyen: (1) la modificación de los medios y límites temporales del examen, así como el contenido de las pruebas existentes, y (2) la evaluación clínica individualizada que combina las calificaciones de la prueba con otras fuentes de datos de la historia del sujeto, entrevistas y el juicio de observadores cotidianos adecuadamente informados, como los maestros (AERA, APA, NCME, 1985, capítulo 13; Bailey y Wolery, 1989; Barnett, 1983; Culbertson y Willis, 1993; Eyde, Nester, Heaton y Nelson, 1994; Scarpati, 1991; Sherman y Robinson, 1982).

Los esfuerzos por establecer normas separadas para personas con discapacidades físicas específicas o para diseñar pruebas especiales para esos grupos suelen verse obstaculizados por el pequeño número de casos disponibles. Esta limitación se aplica en particular a la baja incidencia y múltiples discapacidades y al uso de pruebas en contextos (graduados y admisión a escuelas profesionales) que incluyen muestras especialmente seleccionadas. No obstante, continúa la investigación sobre el desempeño de personas con diferentes discapacidades físicas a las que se han aplicado estándares o versiones especialmente adaptadas de varias pruebas.

El Servicio de Examinación Educativa realizó un ambicioso programa de investigación mediante el uso de versiones estándares y especiales del SAT de la Junta Universitaria y el GRE o Prueba General con cuatro clases de solicitantes discapacitados: con deficiencias auditivas, con deficiencias visuales, con problemas de aprendizaje y con deficiencias físicas (Willingham *et al.*, 1988). Las características psicométricas investigadas incluyeron la confiabilidad, el funcionamiento diferencial de los reactivos, la

[16] En Nester (1994) se encuentra una revisión de las implicaciones de la ADA en la examinación psicológica. Los temas psicométricos y evaluativos generados por esta ley se analizan exhaustivamente en una declaración de la División de Evaluación, Medición y Estadística de la Asociación Estadounidense de Psicología que se publicó en el número de enero de 1993 de *The Score*, el boletín de la división.

[17] En Wachs y Sheehan (1988) puede encontrar mayor información sobre los patrones de desarrollo en los niños pequeños con discapacidad y sobre las consideraciones de procedimiento para su evaluación.

estructura factorial y otros índices de validez relacionados con los niveles de ejecución y la capacidad predictiva; también se estudiaron el contenido, el tiempo y los ajustes de la prueba. En general, los resultados muestran que, en casi todos los aspectos e incluso en el significado de las calificaciones, las adaptaciones del procedimiento son comparables a la examinación estándar; sin embargo, la predicción del rendimiento académico a partir de las calificaciones, o de las calificaciones previas, no es tan exacta para las personas discapacitadas como para los demás, y existen interrogantes sobre la estructura factorial y el funcionamiento de los reactivos de algunas de las adaptaciones de las pruebas (R. E. Bennett, Rock y Novatkoski, 1989; Rock, Bennett y Jirele, 1988; Willingham, 1988). Además, se ha descubierto que los límites temporales en las versiones no estándares son relativamente más indulgentes, un descubrimiento que apoya la práctica controversial de "disminuir" las calificaciones obtenidas con estas versiones, lo que indica la impostergable necesidad de desarrollar de manera empírica límites temporales comparables (véase por ejemplo, Wainer, 1993a, pp. 9–10).

Persisten aún muchos problemas éticos y psicométricos concernientes a la examinación de personas discapacitadas y, aunque se requiere de mayor investigación, también es necesario reconocer que algunos de los problemas pueden ser irresolubles toda vez que se derivan del hecho de que *cada* persona presenta una configuración única de tipos y grados de capacidades, discapacidades y características personales. No obstante, para fines prácticos, existe ahora mayor conciencia y sensibilidad hacia las necesidades de los discapacitados, así como una mejor orientación sobre la manera de aplicarles el examen. Además, los avances en la elaboración de instrumentos (como los sintetizadores de voz que simulan el habla y otros dispositivos electrónicos controlados por computadora) prometen innovaciones en la examinación, muchas de las cuales pueden ser valiosas en esta área (véase, por ejemplo, Educational Testing Service, 1992; Wilson, 1991).

En las siguientes secciones consideraremos los problemas y procedimientos de la examinación especial en relación con las tres categorías principales de discapacidad física, a saber, las deficiencias auditiva, visual y motora.

Deficiencias auditivas.[18] En general, los niños con deficiencias auditivas están en desventaja al aplicarles pruebas verbales por su retardo en el desarrollo del lenguaje, aunque el contenido verbal se les presente visualmente. Entre más temprano inicie la sordera en los niños pequeños, más grave será la deficiencia. Por fortuna, los avances recientes en la evaluación del funcionamiento auditivo han hecho posible diagnosticar con precisión las pérdidas de audición e iniciar los procedimientos de rehabilitación en los primeros meses de vida (Shah y Boyden, 1991).

La examinación de niños sordos fue el principal objetivo del desarrollo de algunas de las primeras escalas de ejecución, como la Escala de Ejecución de Pintner-Paterson (*Pintner-Paterson Performance Scale*) y la Escala Ejecutiva de Arthur (*Arthur Performance Scale*). Con frecuencia se emplean adaptaciones especiales de las escalas de Wechsler para examinar a sordos e incluso es posible aplicar la mayor parte de las pruebas verbales si las preguntas orales se escriben en tarjetas. Se han elaborado varios procedimientos para comunicar las instrucciones de las pruebas de ejecución (véase, por ejemplo, Sattler, 1988, 1992); de hecho, la Escala de Ejecución del WISC-R es el

[18] Para una revisión de los temas y consideraciones correspondientes a la evaluación de niños con deficiencias auditivas, véase Bradley-Johnson y Evans (1991), Mullen (1992), y Sullivan y Burley (1990).

test de inteligencia que más se aplica en los Estados Unidos a los niños con deficiencias auditivas. Sin embargo, dadas las modificaciones de los procedimientos estandarizados de examinación, no es posible suponer que la confiabilidad, la validez y las normas permanecerán inalteradas. Aun así, gracias al uso intensivo de las escalas de Wechsler en personas con problemas auditivos se cuenta con abundante información sobre sus propiedades psicométricas cuando se aplican a muestras de individuos con problemas de audición (véase, por ejemplo, Braden, 1985; Maller y Braden, 1993; Sullivan y Schulte, 1992). En general, esos estudios indican que existen semejanzas sustanciales en la estructura factorial de las escalas y en la validez predictiva y de constructo de la Escala de Ejecución para los niños con y sin deficiencias auditivas.

Las pruebas que hemos citado hasta ahora se estandarizaron en personas con audición normal. Muchos investigadores han llegado a la conclusión de que cuando los niveles de ejecución son comparables a los de las personas con audición normal, como es el caso de las escalas de ejecución de Wechsler, no hay necesidad de tener normas separadas para los individuos con deficiencias auditivas. Al mismo tiempo, las normas obtenidas con niños sordos resultan útiles en diversas situaciones pertinentes a su desarrollo educativo. Para satisfacer esta necesidad se han realizado esfuerzos por establecer normas especiales para las pruebas existentes, como lo ilustra la estandarización de la Escala de Ejecución para Niños Sordos del WISC-R (*Performance Scale for Deaf Children;* R.J. Anderson y Sisco, 1977).

A un nivel más elemental, el Test de Aptitud para el Aprendizaje de Hiskey-Nebraska (*Hiskey-Nebraska Test of Learning Aptitude*, Hiskey, 1966) se desarrolló y estandarizó en niños sordos y parcialmente sordos. Se trata de una prueba individual adecuada para edades de tres a 17 años, en la que se eliminó la velocidad porque es difícil transmitir la idea de rapidez a un niño sordo. Se hizo un intento por probar un arreglo de funciones intelectuales mayor al cubierto por la mayor parte de las pruebas de ejecución. En la prueba se utilizan la pantomima y ejercicios de práctica para comunicar las instrucciones, así como reactivos intrínsecamente interesantes que establecen el *rapport*. Los reactivos se eligieron haciendo referencia especial a las limitaciones de los niños sordos y la selección final se basó ante todo en el criterio de diferenciación por edad. Se derivaron normas por separado de 1 079 niños sordos y 1 074 con audición normal. En el manual se proporcionan instrucciones paralelas para examinar a ambos grupos de niños y una revisión de las prácticas que resulta conveniente seguir al examinar a niños sordos. Aunque las normas son anticuadas, el Hiskey-Nebraska presenta evidencia adecuada sobre su confiabilidad y validez y todavía se considera una de las mejores pruebas para niños con deficiencias auditivas (Sullivan y Burley, 1990).

Durante las últimas cinco décadas se ha observado una sorprendente evolución del conocimiento sobre las consecuencias intelectuales de la sordera. Braden (1994) analiza gran parte de esa historia en una cuidadosa revisión de más de 200 estudios de sordos que incluyen a más de 170 000 participantes. El trabajo describe muchos de los descubrimientos interesantes, como el hecho de que los niños sordos que tienen padres sordos obtienen calificaciones por encima de las normas para niños sin discapacidades en las pruebas de ejecución. Aunque todavía no se comprenden las razones de estos descubrimientos, resulta indudable que la sordera es una variable mucho más compleja de lo que se pensaba. La etiología, el grado, la edad de inicio y la detección de la pérdida auditiva, así como el modo de comunicación, la colocación educativa, el estado auditivo de los padres y la existencia de discapacidades adicionales son factores

que interactúan y contribuyen a la ejecución cognoscitiva diferencial de los individuos con deficiencias auditivas.

Deficiencias visuales.[19] La evaluación de los ciegos supone problemas muy diferentes a los que se enfrentan con los sordos; por ejemplo, es más sencillo adaptar pruebas orales para las personas ciegas, pero es más difícil aplicarles las de ejecución. Además de la presentación oral por parte del experimentador, pueden utilizarse otras técnicas, como el uso de grabaciones. Algunos instrumentos cuentan con formatos con caracteres grandes o en braille, como las Test de Evaluación Escolar de la Junta Universitaria (*College Board Scholastic Assessment Test*, SAT), aunque la aplicación de esta última se ve restringida por el gran volumen de materiales impresos en braille en comparación con los impresos en tinta, por la tasa más lenta de lectura en braille y por el número de ciegos que no leen braille con facilidad. Las respuestas de los examinados pueden registrarse en braille o en un tablero, aunque también se dispone de hojas o tarjetas de respuesta preparadas en relieve para usar con reactivos de falso o verdadero, de opción múltiple y otros de selección de respuesta. Por supuesto, en muchas pruebas de aplicación individual es posible obtener respuestas orales o mediante gestos.

Entre los primeros ejemplos de los tests de inteligencia general adaptados para los ciegos está el de Binet. La primera revisión Hayes-Binet para examinar a los ciegos se basó en la versión de 1916 de la escala Stanford-Binet. En 1942 se preparó la Hayes-Binet Provisional[20] a partir de la versión de 1937 del Stanford-Binet (Hayes, 1942, 1943). La adaptación más reciente, comparable a la forma L-M del Stanford-Binet, es el Test de Inteligencia para Ciegos de Perkins-Binet (*Perkins-Binet Tests of Intelligence for the Blind*), un instrumento estandarizado que cuenta con formas separadas para niños ciegos y débiles visuales (C. J. Davis, 1980).

Las escalas de Wechsler también se adaptaron para examinar a ciegos. En este caso las adaptaciones consisten principalmente en el uso de pruebas verbales, la omisión de las pruebas de ejecución, y la sustitución de algunos reactivos inadecuados por otros. En general, los estudios de niños ciegos o débiles visuales indican que estas condiciones pueden tener un impacto negativo sobre el desarrollo cognoscitivo, incluso en el área verbal, por las limitaciones que imponen sobre la amplitud y variedad de sus experiencias. En distintos estudios se ha encontrado un patrón similar entre los perfiles Wechsler de niños con deficiencias visuales; los resultados plantean que, en su caso, la composición factorial de las tareas es diferente a la de los niños con visión normal. Aunque los CI globales no pueden considerarse medidas exactas del funcionamiento cognoscitivo total de los niños con deficiencias visuales, en manos de un examinador experto las escalas de Wechsler pueden proporcionar información de diagnóstico útil sobre los puntos fuertes y las debilidades de esos niños (Groenveld y Jan, 1992).

Son muy pocos los instrumentos desarrollados específicamente para personas con deficiencias visuales, quizás el ejemplo mejor conocido sea el Test de Aptitud para el Aprendizaje de los Ciegos (*Blind Learning Aptitude Test*, BLAT, Newland, 1979), una prueba de aplicación individual que incorpora reactivos adaptados de otros instru-

[19] Véase Bradley-Johnson (1994), Fewell (1991), M. S. Moore y McLaughlin (1992) y Orlansky (1988) para revisiones de temas y procedimientos en la evaluación de niños con deficiencias visuales.

[20] Debido a la naturaleza tentativa de su estandarización, esta revisión se diseñó originalmente como una edición provisional por lo que llegó a conocerse con ese nombre.

mentos (como las Matrices Progresivas de Raven) y otros reactivos no verbales, y los presenta en relieve. La prueba pone más énfasis en el proceso de aprendizaje que en sus productos para no poner en desventaja al niño ciego. Los datos normativos, aunque anticuados, se comparan favorablemente con los disponibles para instrumentos diseñados para poblaciones especiales. Por otro lado, la información sobre la confiabilidad y la validez es insuficiente y requiere de mayor investigación. No obstante, el BLAT puede ser un componente útil, junto con otras pruebas verbales, para la evaluación de los niños ciegos de edad escolar.

Un ejemplo más reciente, aunque aún en desarrollo, es la tentativa de un grupo de investigadores holandeses para medir la inteligencia de niños con deficiencias visuales. El Test de Inteligencia para Niños con Deficiencias Visuales (*Intelligence Test for Visually Impaired Children*, ITVIC, Dekker, Drenth, Zaal y Koole, 1990) incluye versiones táctiles de tareas como el Diseño con Cubos en una batería que incluye varios subtests verbales y no verbales.[21] Aunque los estudios iniciales han mostrado resultados prometedores en cuanto a la validez del instrumento, es necesario realizar más investigaciones sobre el ITVIC con un mayor número de sujetos (Dekker, 1993; Dekker, Drenth y Zaal, 1991; Dekker y Koole, 1992).

Como en el caso de las otras condiciones especiales que estudiamos en este capítulo, las deficiencias visuales se presentan en distintos grados y muy a menudo en combinación con otros problemas, por lo que la decisión de usar pruebas estandarizadas, adaptaciones o pruebas especialmente diseñadas para los invidentes depende de los objetivos de la evaluación y de las características peculiares de las personas en cuestión. En general, los usuarios deben recordar que no es posible suponer que las modificaciones de las pruebas, como las presentaciones táctiles de los diseños visuales o la extensión de los límites temporales, miden los mismos constructos que las versiones originales.

Deficiencias motoras.[22] Aunque por lo general pueden recibir la estimulación auditiva y visual, las personas con discapacidades ortopédicas pueden presentar trastornos motores tan graves que resulta imposible obtener respuestas orales o escritas. El intento por manipular tableros u otros materiales de ejecución puede enfrentar dificultades; además, trabajar con un límite temporal o en un entorno extraño a menudo aumenta las alteraciones motrices del discapacitado ortopédico, cuya mayor susceptibilidad a la fatiga hace necesario acortar las sesiones de prueba.

Algunas de las discapacidades motoras más graves se encuentran entre las personas con parálisis cerebral, aunque en la revisión de esos casos es común el uso de tests comunes de inteligencia como la escala Stanford-Binet. En esos estudios, de los que generalmente se excluye a los más discapacitados por considerar que no pueden ser examinados, es frecuente que se hagan ajustes informales a los procedimientos del examen para adaptar el instrumento a las capacidades de respuesta del niño. Desde luego, ambos procedimientos son provisionales.

---

[21] La Escala de Inteligencia Táctil es una prueba similar desarrollada y normalizada en adultos ciegos durante los cincuenta e inicios de los sesenta (Shurrager y Shurrager, 1964). Consta de seis subtests de ejecución modelados de acuerdo con la Escala de Inteligencia Wechsler-Bellevue, a saber, Símbolos en Dígitos, Ensamble de Objetos, Diseño con Cubos, Completamiento de Objetos, Tablero de Patrones y Cuentas Aritméticas.

[22] En H. G. Williams (1991) puede encontrarse una revisión de las medidas útiles para evaluar las funciones motoras gruesas en los niños pequeños. C. Robinson y Fieber (1988) describen una aproximación orientada al proceso para la evaluación de niños pequeños que emplea las tareas sensoriomotoras y preoperacionales de Piaget.

Un planteamiento más satisfactorio consiste en preparar instrumentos de prueba que sean adecuados incluso para los individuos más discapacitados. Pero aunque ya se dispone de algunas pruebas especialmente diseñadas con este propósito o con adaptaciones de los instrumentos ya existentes, a menudo son escasos los datos sobre su normatividad y su validez. Algunas de las pruebas que veremos en la siguiente sección, diseñadas originalmente para la examinación transcultural, han demostrado ser aplicables a los discapacitados; por ejemplo, se cuenta con adaptaciones de la Escala Internacional de Ejecución de Leiter (*Leiter International Performance Scale*) y de los Laberintos de Porteus que pueden aplicarse a niños con parálisis cerebral (Allen y Collins, 1955; Arnold, 1951). En ambos casos, el examinador manipula los materiales de la prueba y el examinado responde con movimientos de cabeza. También las Matrices Progresivas de Raven resultan una herramienta útil para este propósito y parecen ser en especial apropiadas para individuos con discapacidades ortopédicas, ya que se aplican sin límite de tiempo y el individuo puede responder oralmente, por escrito, con ademanes o con gestos como una inclinación de cabeza. Pese a la flexibilidad y simplicidad del indicador de respuesta de las diversas formas de la prueba, no sólo cubre un amplio rango de dificultad, sino que también proporciona un techo más bien elevado. Diversos estudios en individuos con parálisis cerebral y otros trastornos motores han informado del uso exitoso de esta prueba (véase, por ejemplo, Capitani, Sala y Marchitti, 1994).

Otras pruebas que permiten el uso de una respuesta simple como señalar son las *pruebas pictóricas de vocabulario*, las cuales proporcionan una medida rápida del "uso" del vocabulario que las hace adecuadas para personas que no pueden vocalizar (como quienes sufren parálisis cerebral) y para los sordos. Como su aplicación es sencilla y pueden completarse aproximadamente en 15 minutos, también resultan útiles como instrumento de detección rápida en situaciones en que no es factible aplicar un test comprensivo de inteligencia.

El Test de Vocabulario con Ilustraciones de Peabody es un ejemplo de esos instrumentos. Su revisión actual (PPVT-R, Dunn y Dunn, 1981) consta de una serie de 175 láminas, cada una de las cuales contiene cuatro ilustraciones. Conforme se presenta cada lámina, el examinador dice una palabra de estímulo y el examinado responde señalando o designando de otra manera el dibujo que mejor ilustra su significado. Aunque la prueba entera abarca del nivel preescolar al adulto, a cada individuo sólo se le presentan los reactivos adecuados para su nivel de ejecución, determinado por una carrera de aciertos en un extremo y de desaciertos en el otro. Las puntuaciones no procesadas o directas pueden convertirse en calificaciones estándares ($M = 100$, $DE = 15$), rangos percentilares y estaninas. Esas calificaciones derivadas se grafican en una tabla con bandas de confianza que cubren ±1 EEM (error estándar de medición), además de proporcionar equivalentes por edad. El PPVT-R no tiene límite de tiempo, pero requiere de 10 a 20 minutos, y está disponible en dos formas paralelas, cada una de las cuales utiliza un conjunto distinto de láminas y diferentes palabras estímulo.

El PPVT-R se estandarizó en una muestra nacional estadounidense de 4 200 niños y jóvenes de entre dos y medio y 18 años y 828 adultos de 19 a 40 años, y sus propiedades psicométricas son muy satisfactorias (para una revisión, véase McCallum, 1985, y Wiig, 1985). Los coeficientes de confiabilidad, encontrados mediante procedimientos de consistencia interna, formas alternas y retest, van de moderados a altos. Buena parte de la evidencia a favor de la validez del PPVT-R se fundamenta en una sólida

base de investigación realizada sobre el instrumento, que tiene una correlación mediana de .70. En una revisión de más de 300 estudios que utilizaron el PPVT se encontró que su correlación con otras pruebas de vocabulario es elevada, moderada con tests de inteligencia verbal y aptitud académica y prometedora con las pruebas de ejecución y aprovechamiento académico. Las correlaciones fueron similares en diversas poblaciones, incluyendo a grupos en desventaja económica y muestras de personas con discapacidades y retardo mental. Las calificaciones del PPVT reflejan en parte el grado de asimilación cultural y de exposición al inglés estadounidense común del examinado.

Los estudios que utilizan el PPVT-R indican que también presenta una elevada correlación con otras medidas de comprensión verbal (véase, por ejemplo, Elliott, 1990b, p. 235). Un estudio particularmente interesante del PPVT-R, que utiliza un modelo de ecuación estructural del orden de adquisición de palabras, ofrece un apoyo importante a la validez de constructo del instrumento (L. T. Miller y Lee, 1993). Para finales de los noventa se espera la publicación de la tercera edición del Test de Vocabulario con Ilustraciones de Peabody, ya que se encuentra en desarrollo.

Los *Tests de Clasificación Pictórica*, como la Escala de Madurez Mental de Columbia (*Columbia Mental Maturity Scale*, CMMS, Burgemeister, Blum y Lorge, 1972) utilizan procedimientos similares para aplicar el instrumento. La escala mencionada (que se formuló originalmente para niños con parálisis cerebral) comprende 92 reactivos, cada uno de los cuales consta de un conjunto de tres, cuatro o cinco dibujos coloreados impresos en una tarjeta grande. Al examinado se le pide que identifique el dibujo que no se relaciona con los otros y que indique su elección con una señal o con una inclinación de cabeza (véase la figura 9.3). La muestra de estandarización para la CMMS estuvo constituida por 2 600 niños entre los tres años con seis meses y los nueve años con 11 meses de edad y fue representativa de la población estadounidense de 1960. La confiabilidad de división por mitades y la de retest fluctuó entre .84 y .91; además, se descubrió una correlación de .67 con el CI del Stanford-Binet en un grupo de 52 preescolares y niños de primer grado y sus correlaciones con las calificaciones de pruebas de aprovechamiento en muestras de primer y segundo grados se cargan principalmente arriba a .40 y en la parte inferior a .60. Aunque se cuenta con muchos datos sobre la validez y aplicabilidad de la versión anterior y la actual de la CMMS para varios grupos de individuos discapacitados (véase *Tests in Print II, III* and *IV*), sus normas anticuadas y la estrecha gama de habilidades que evalúa restringen considerablemente la conveniencia de aplicar dicho instrumento.

EXÁMENES MULTICULTURALES

El problema. Desde la década de los cincuenta se considera con atención el examen de personas con antecedentes culturales distintos. En muchas partes del mundo se necesitan pruebas que faciliten la óptima utilización de los recursos humanos de las naciones en desarrollo, donde la rápida expansión de las instalaciones educativas requiere la aplicación de exámenes tanto para propósitos de admisión como de consejería. Al aumentar la industrialización también crece la demanda de pruebas que ayuden en la selección y distribución del personal, particularmente en los campos del procesamiento de información, mecánico y profesional.

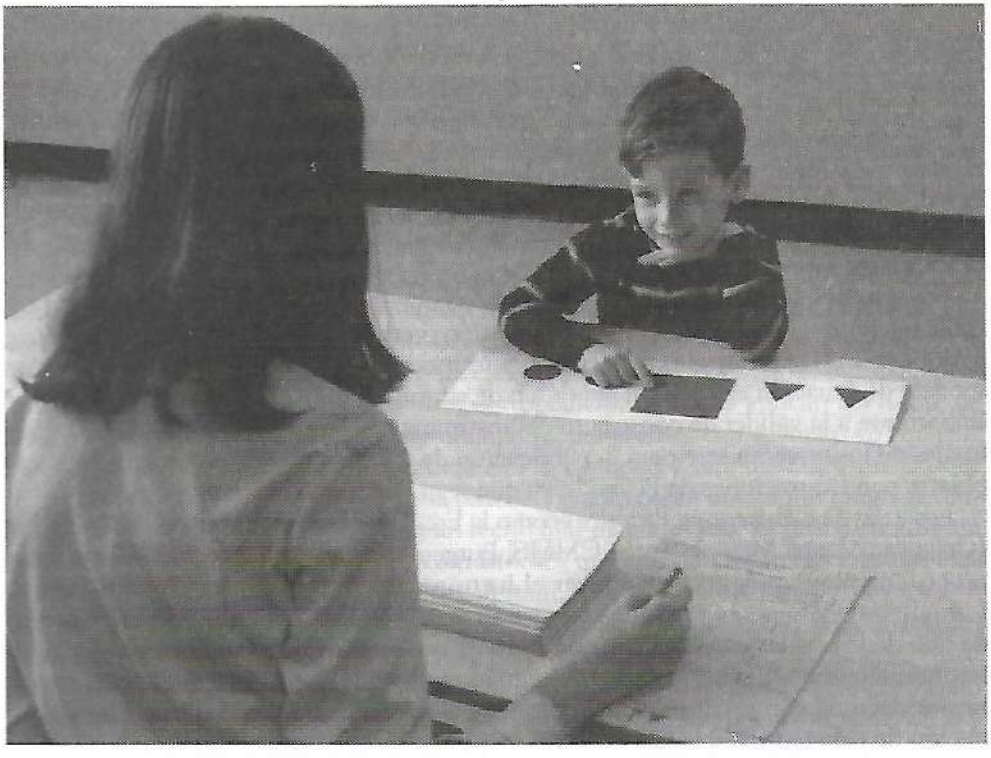


Figura 9.3. La examinadora aplica al niño la Escala de Madurez Mental Columbia.

(Tomado de *Columbia Mental Maturity Scale: Guide for administering and interpreting*. Burgemeister *et al.*, 1972, p. 11. Copyright © 1972 por Harcourt Brace Jovanovich, Inc. Reproducido con autorización).

En los Estados Unidos los problemas prácticos de las pruebas multiculturales[23] se han asociado principalmente con la existencia de culturas minoritarias dentro de la cultura dominante. Existe una difundida preocupación por la aplicabilidad de las pruebas disponibles para las personas con desventajas culturales, por lo que debe reconocerse que la desventaja cultural es un concepto relativo. De manera objetiva, sólo existen diferencias entre dos culturas o subculturas cualesquiera, pues cada una refuerza el desarrollo de un comportamiento adaptado a sus valores y demandas. Cuando un individuo debe ajustarse y competir dentro de una cultura o subcultura diferente de la suya, la diferencia puede convertirse en desventaja o, por el contrario, en ventaja cultural.

Aunque el interés por los exámenes transculturales se ha visto estimulado por los recientes avances sociales y políticos, el problema se reconoció al menos desde 1910. Algunas de las primeras pruebas transculturales se destinaron a examinar a las grandes olas de inmigrantes que llegaron a los Estados Unidos durante las primeras décadas del siglo XX (Knox, 1914). Otras tuvieron su origen en la investigación básica sobre las

[23] Las "pruebas multiculturales" también se conocen como pruebas transculturales.

habilidades comparativas de grupos culturales relativamente aislados, que a menudo tenían poco o ningún contacto con las sociedades occidentales tecnológicamente avanzadas en cuyo marco se elaboró la mayor parte de las pruebas psicológicas.[24]

Tradicionalmente, las pruebas transculturales pretenden descartar uno o más parámetros en los que varían las culturas, uno de los cuales es el *idioma*. Si los grupos culturales que van a examinarse hablan idiomas distintos, se hace necesario desarrollar pruebas que no requieran del lenguaje de parte del examinador o de los examinados. Otro ejemplo fue la supresión de la *lectura* cuando los antecedentes educativos diferían considerablemente y dominaba el analfabetismo; en este caso, no se eliminó el lenguaje oral porque las pruebas se diseñaron para personas que hablaban el mismo idioma. La *velocidad* es otro parámetro en el que difieren las culturas o subculturas, es decir, las culturas nacionales, grupos étnicos minoritarios y las subculturas urbana y rural no sólo difieren en el ritmo de su vida cotidiana, sino también en la motivación por apresurarse y el valor que asignan a la ejecución rápida (véase por ejemplo, Klineberg, 1928; R. R. Knapp, 1960; M. Womer, 1972), razón por la cual las pruebas transculturales a menudo, aunque no siempre, han intentado eliminar la influencia de la velocidad al establecer límites temporales amplios y al no ofrecer premios por una ejecución rápida.

Otros parámetros en los que difieren las culturas tienen que ver con el *contenido de la prueba;* por ejemplo, algunas pruebas sin lenguaje y sin lectura demandan reactivos de información que son específicos de ciertas culturas. Así, pueden exigir que el examinado entienda la función de objetos como un violín, un sello postal, una pistola, una navaja, un teléfono, un piano o un espejo, por lo que las personas educadas en ciertas culturas pueden carecer de la experiencia que les permita responder de manera correcta a esos reactivos. Las primeras pruebas "libres de cultura" se desarrollaron principalmente para controlar este parámetro cultural. Luego de un breve examen de las pruebas diseñadas para eliminar uno o más de los parámetros mencionados regresaremos al análisis de aproximaciones alternativas a la examinación transcultural.

Instrumentos tradicionales clásicos.[25] En su esfuerzo por desarrollar pruebas que pudieran aplicarse a culturas distintas, los psicómetras siguieron diversos procedimientos, algunos de los cuales se ejemplifican en las pruebas que estudiamos en esta sección. La Escala Internacional de Ejecución de Leiter, Revisada (Roid y Miller, 1997) es un instrumento individual de ejecución que se publicó originalmente en 1940. Luego se utilizó durante varios años con diferentes grupos étnicos en Hawai más tarde, Porteus la aplicó a varios grupos africanos y otros investigadores a otros grupos nacionales. Una revisión publicada en 1948 se basó en la evaluación posterior de niños, estudiantes de secundaria y reclutas estadounidenses durante la Segunda Guerra Mundial. La edición de 1997 se basa en muestras de más de 2 000 individuos estadounidenses normales y atípicos de entre dos y 20 años de edad. Un rasgo distintivo de la Leiter, que luego adoptaron otros instrumentos, es la eliminación casi absoluta de las instrucciones verbales. Cada prueba empieza con una tarea muy sencilla del tipo que

[24] En Anastasi (1954, capítulo 10) se encuentran ejemplos de las primeras pruebas.

[25] En Naglieri y Prewett (1990) pueden encontrarse revisiones de las medidas no verbales que analizamos en esta sección y en la anterior, así como de otras pruebas de esa categoría.

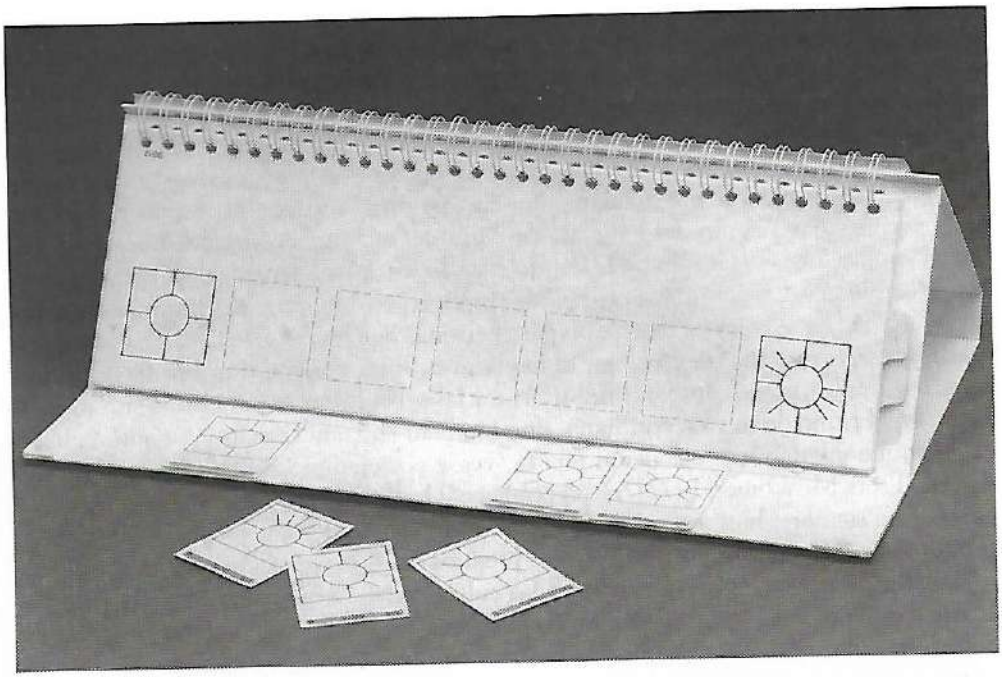


Figura 9.4. Materiales clásicos en el uso de la Escala Internacional de Ejecución de Leiter-Revisada. La tarea parcialmente completada que se ilustra pertenece a la Prueba de Orden de Secuencias. Requiere que el examinado elija cinco de seis tarjetas de respuesta y que las coloque correctamente en la gaveta.

(Cortesía de la Stoelting Company.)

puede encontrarse en el resto del instrumento y se considera que la comprensión de las tareas (que se aplican individualmente y sin límites de tiempo) forma parte de la prueba. Como se ilustra en la figura 9.4, todos los materiales de estímulo gráficos se presentan en caballetes y el examinado responde eligiendo las tarjetas con las imágenes más apropiadas que coloca en la bandeja de respuestas.

La escala de Leiter se diseñó para abarcar una amplia variedad de funciones, similares a las encontradas en las escalas verbales. En su forma actual, la cobertura ha aumentado significativamente y ahora incluye cuatro áreas: Razonamiento, Visualización, Atención y Memoria. Entre las tareas incluidas en las distintas edades en las áreas de razonamiento y visualización se encuentran: analogías de diseños, completamiento de formas, igualación, ordenamiento de secuencias (ilustrada en la figura 9.4), clasificación, doblado de papel y rotación de figuras. Las pruebas de las áreas de Atención y Memoria incluyen medidas de atención (sostenida y dividida) así como diversas tareas de memoria inmediata y demorada. Como era de esperarse, la escala revisada de Leiter se ha actualizado considerablemente y es más sofisticada que las versiones anteriores respecto de sus características psicométricas; por ejemplo, para calibrar los grados de dificultad se utilizó la teoría de respuesta al ítem y las calificaciones ya no se expresan

como razones de CI. Además, al contar con normas representativas contemporáneas y al aumentar el contenido se espera que la utilidad de la escala sea mayor. El nuevo manual incluye información sobre la evidencia de varios tipos de confiabilidad y validez.

Las Matrices Progresivas de Raven (MPR) se diseñaron principalmente como una medición del factor g de Spearman o inteligencia general (J. Raven, 1983; Raven, Raven y Court, 1995). De acuerdo con el análisis teórico que Spearman hizo de g, esta prueba requiere principalmente la deducción de relaciones entre reactivos abstractos, que consisten en un conjunto de matrices o arreglos de diseños en renglones y columnas de los cuales se eliminó una parte. La tarea consiste en elegir la parte faltante entre las alternativas proporcionadas. Los reactivos más sencillos requieren una discriminación precisa; los más difíciles incluyen analogías, permutaciones, alternación de patrones y otras relaciones lógicas. En la figura 9.5 se presentan dos reactivos de las Matrices Progresivas Estándares. La prueba no suele incluir límites temporales y puede aplicarse de manera individual o colectiva, además de que requiere instrucciones orales muy sencillas.

Se dispone de tres formas de las MPR que difieren en su grado de dificultad. Las Matrices Progresivas Estándar (MPE, edición de 1996) es la forma adecuada para individuos promedio entre seis y 80 años. Una forma más sencilla, las Matrices Progresivas de Colores (MPC, edición de 1990) es conveniente para niños más pequeños y para grupos especiales que por diversas razones no pueden examinarse con las MPE. En el caso de las MPC se cuenta con normas para niños de cinco y medio a 11 y medio años, para muestras de personas no retardadas de 60 a 89 años y para adultos con retardo mental. La tercera forma, las Matrices Progresivas Avanzadas (MPA, edición de 1994), se destinó a adolescentes y adultos superiores al promedio.

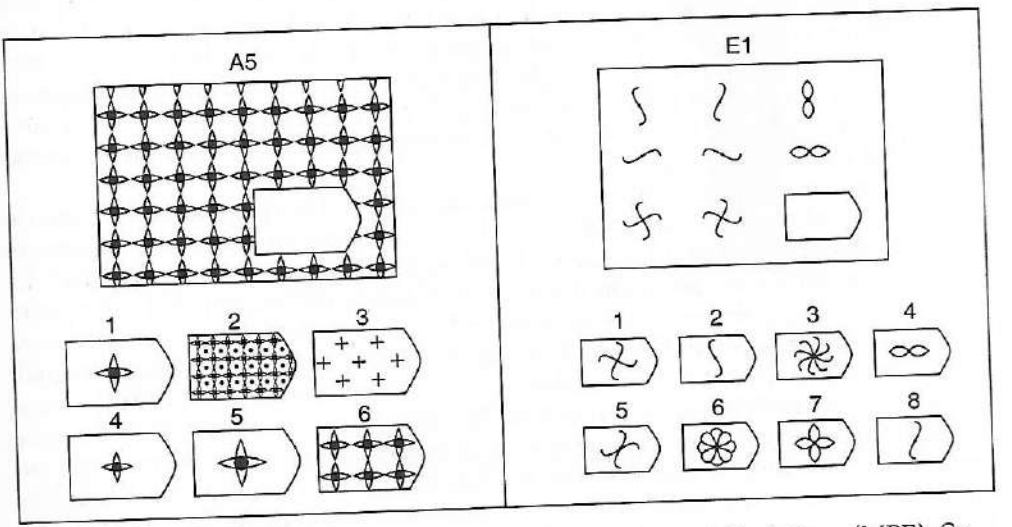


Figura 9.5. Reactivos clásicos de las Matrices Progresivas Estándares (MPE). Se muestra un reactivo sencillo (A5) y uno difícil (E1)

(Reproducido con permiso de J. C. Raven Ltd.)

El manual para todos los niveles de las Matrices Progresivas de Raven está disponible en secciones que pueden comprarse por separado o en cualquier combinación deseada y que pueden unirse en una sola carpeta. La primera sección ofrece una visión general y se actualizó en 1995, las otras se publicaron entre 1990 y 1996 e incluyen los manuales específicos para cada uno de los tres niveles de la prueba. También forman parte del paquete de las Raven los manuales para dos pruebas de vocabulario estandarizadas que pueden emplearse en combinación con las MPR. La última sección del manual compendia la investigación adicional sobre la confiabilidad, validez y normas complementarias obtenidas en diversos países y con poblaciones especiales (Court y Raven, 1995). También se cuenta con varios apéndices de investigación que incluyen datos de estandarización obtenidos en Inglaterra, información normativa adquirida en los Estados Unidos, Irlanda y Alemania, y una bibliografía anotada de más de 2 000 estudios.[26]

Aunque la investigación publicada sobre las MPR es considerable, los estudios son dispersos y heterogéneos, ya que se diseñaron con propósitos muy diversos. Los autores de la prueba recomiendan que los usuarios potenciales o certificados identifiquen los estudios y poblaciones más relevantes a sus intereses, pero advierten que existen diferencias considerables en la metodología, tamaño de las muestras y calidad técnica de esos estudios.

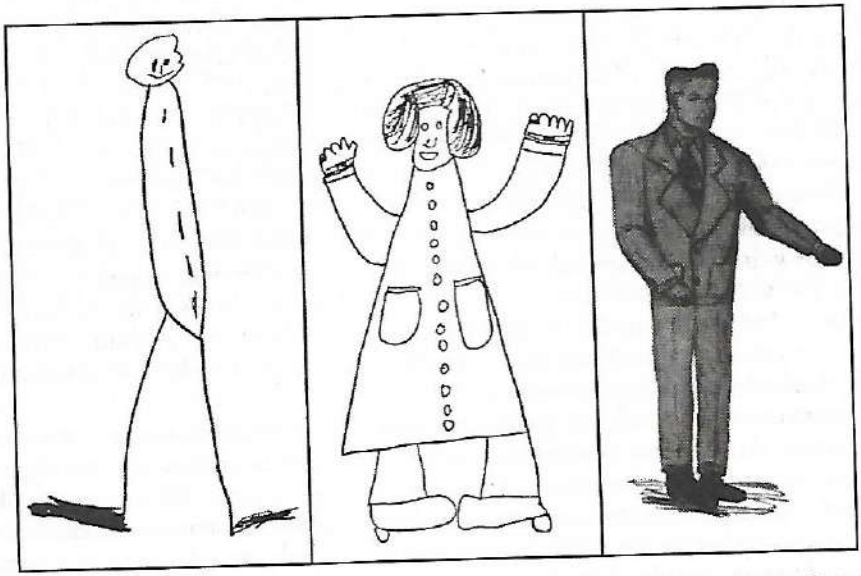
En general, la confiabilidad del retest en grupos de niños mayores y adultos de edad moderadamente homogénea fluctúa entre .70 y .90. Sin embargo, en los rangos inferiores de calificación la confiabilidad cae muy por debajo de estas cifras. Los coeficientes de consistencia interna se encuentran principalmente entre .80 y .90; las correlaciones con las pruebas de inteligencia verbales y de ejecución caen entre .40 y .75, tienden a ser mayores en las pruebas de ejecución que en las verbales. Los estudios con personas mentalmente retardadas y con diferentes grupos ocupacionales y educativos indican una buena validez concurrente. Los coeficientes de validez predictiva de criterio académico son algo menores a lo que es común con las pruebas de inteligencia verbal. Varios análisis factoriales sugieren que la prueba de Matrices Progresivas está muy cargada con un factor común a la mayor parte de las medidas de inteligencia (identificado por muchos psicólogos con la *g* de Spearman), pero la ejecución también se ve afectada por la aptitud espacial, el razonamiento inductivo, la exactitud perceptual y otros factores grupales.

Una aproximación diferente a la evaluación no verbal la ejemplifica la Prueba de Dibujo de un Hombre (*Draw-a-Man Test*) de Goodenough, en la que simplemente se pide al examinado "haz el dibujo de un hombre; haz el mejor dibujo que puedas". La prueba se utilizó sin modificación desde su estandarización original en 1926 y hasta 1963, cuando se publicó una extensión revisada con el título de Prueba de Dibujo de Goodenough-Harris (D. B. Harris, 1963). En la revisión, como en la versión original, se hace énfasis en la exactitud de la observación del niño y en el desarrollo del pensamiento conceptual más que en la habilidad artística. Se da crédito por la inclusión de partes individuales del cuerpo, detalles de la vestimenta, proporción, perspectiva y características similares. Se seleccionó un total de 73 reactivos calificables sobre la base de la diferenciación por edad, relación con las calificaciones totales en la prueba y relación con las calificaciones del test colectivo de inteligencia.

[26] Es posible obtener la última versión en disco o en copia con J. H. Court, en el domicilio del registro de la editorial de las MPR.

En las escalas revisadas también se pide a los examinados que hagan el dibujo de una mujer y uno de sí mismos. La escala de la mujer se califica de manera similar a la del hombre. La escala de sí mismo se desarrolló como un test proyectivo de personalidad, aunque los resultados no son prometedores.[27] Las puntuaciones obtenidas en cada escala se transforman en calificaciones estándares con una media de 100 y una *DE* de 15, La figura 9.6 muestra tres dibujos hechos por niños de cinco años con ocho meses, ocho años con ocho meses y 12 años con 11 meses junto con las puntuaciones crudas (o directas) y las calificaciones estándares correspondientes. La confiabilidad de la Prueba de Dibujo de Goodenough-Harris se ha investigado repetidamente mediante distintos procedimientos. Las correlaciones en el retest así como los valores de confiabilidad del calificador y de la división por mitades resultan adecuadas, además, el efecto de la enseñanza de arte sobre las calificaciones es al parecer insignificante (J. A. Dunn, 1967; D. B. Harris, 1963).

Además de los datos del análisis de reactivos obtenidos al desarrollar las escalas, la información relativa a la validez de constructo se encuentra en las correlaciones del instrumento con otros tests de inteligencia, que si bien varían mucho se encuentran



Hombre: Puntuación cruda 7
EC 5:8
Calificación estándar 73

Mujer: Puntuación cruda 31
EC 8:8
Calificación estándar 103

Hombre: Puntuación cruda 66
EC 12:11
Calificación estándar 134

Figura 9.6. Ejemplos de dibujos obtenidos con la Prueba de Dibujo de Goodenough-Harris.

(Cortesía de Dale B. Harris.)

[27] En el capítulo 15 estudiaremos otras aproximaciones proyectivas al uso de los dibujos de la figura humana. Entre éstas se incluye el trabajo de Elizabeth Koppitz con esta técnica, que cubre aspectos cognoscitivos y emocionales.

sobre todo por encima de .50. Para los niños del nivel preescolar, la Prueba de Dibujo de un Hombre muestra una correlación elevada con la aptitud numérica y menor con la velocidad perceptual y la precisión que para los niños de cuarto grado (D. B. Harris, 1963), lo que indica que el instrumento puede medir funciones distintas a diferentes edades. Ambas versiones han sido empleadas en numerosos estudios con grupos étnicos y culturales disímiles, los cuales indican que la prueba depende más de las diferencias en los antecedentes culturales de lo que se había supuesto originalmente; por ejemplo, Dennis (1966) analizó los datos comparativos obtenidos con la prueba en 40 grupos culturales diferentes y descubrió que el promedio de las calificaciones grupales parecía estar estrechamente relacionado con el grado de experiencia con el arte representativo de cada cultura.

Las diferencias culturales en la experiencia ambiental se revelaron también en una bien diseñada investigación comparativa de niños mexicanos y estadounidenses con la prueba de Goodenough-Harris (Laosa, Swartz y Díaz-Guerrero, 1974). En un estudio más reciente de la prueba realizado en Irán, las calificaciones promedio de niños de seis a 13 años fueron algo inferiores a las normas estadounidenses, pero mostraron una buena diferenciación por edad y correlaciones positivas con el nivel socioeconómico y las medidas de aprovechamiento académico (Mehryar, Tashakkori, Yousefi y Khajavi, 1987). Debe agregarse que esos descubrimientos con la prueba de Goodenough-Harris son característicos de los resultados obtenidos con todas las pruebas diseñadas originalmente para ser "libres de cultura" o "culturalmente justas" (Samuda, 1975, capítulo seis).

Se dispone ya de una nueva versión de la Prueba de Dibujo de un Hombre que pretende actualizar y mejorar las cualidades técnicas de la versión de Goodenough-Harris con el nombre de Prueba de Dibujo de una Persona: Un sistema Cuantitativo de Calificación (*Draw a Person: A Quantitative Scoring System, DAP*, Naglieri, 1988). La DAP proporciona normas más recientes y detalladas, procedimientos de aplicación algo distintos y un sistema revisado de calificación que pretende ser menos ambiguo que la Prueba de Dibujo de Goodenough-Harris, y datos normativos de muestras de niños negros e hispanos; empero, a pesar de sus avances, la revisión ha sido criticada por su alcance relativamente limitado y por la falta de evidencia sobre las ventajas del sistema revisado de calificación (Cosden, 1992).

Es necesario reiterar un último punto respecto a los instrumentos que analizamos en esta sección. Aunque inicialmente se desarrollaron para la medición transcultural, algunos han encontrado importantes aplicaciones en los métodos usados por psicólogos clínicos y consejeros, lo mismo para complementar la información obtenida con instrumentos como las escalas de Stanford-Binet y de Wechsler, que como una fuente de datos en la examinación de personas con diversas discapacidades. Este reconocimiento ha tenido como resultado la preparación de una nueva generación de dichas mediciones, una de las cuales, ya en su segunda edición, es el Test de Inteligencia no Verbal (*Test of Nonverbal Intelligence*, TONI-2, L. Brown, Sherbenou y Johnsen, 1990), similar a las Matrices Progresivas de Raven en contenido y amplitud de aplicabilidad (para revisiones del TONI-2, véase K. R. Murphy, 1992, y Watson, 1992). Otros instrumentos importantes de esa categoría se encuentran en estandarización y se publicarán a finales de la década de los noventa.[28]

[28] Por ejemplo, el Test Universal de Inteligencia No Verbal, de B. A. Bracken y R. S. McCallum.

Aproximaciones a la examinación transcultural. En teoría, podemos identificar tres aproximaciones al desarrollo de pruebas para personas educadas en culturas o subculturas diferentes, aunque en la práctica se combinan algunas características de las tres. El primer planteamiento incluye la elección de reactivos que sean comunes a muchas culturas diferentes y la validación de la prueba resultante comparada con los criterios locales en esas distintas culturas. Ésta es la aproximación básica de las pruebas culturalmente justas, aunque su validación repetida en las distintas culturas se ha descuidado o bien se ha realizado de manera inadecuada, lo que nos impide estar seguros de que la prueba está relativamente libre en buena medida de elementos culturalmente restringidos. Más aún, es poco probable que pueda diseñarse una prueba que cumpla a cabalidad estos requisitos en una amplia diversidad de culturas.

Con todo, es necesario contar con técnicas de evaluación multicultural para la investigación básica de algunas cuestiones fundamentales, una de las cuales atañe a la generalidad de los principios y constructos psicológicos derivados en una única cultura (Anastasi, 1958, capítulo 18; Berry *et al.*, 1992; Irvine, 1983; Irvine y Carroll, 1980). Otra cuestión tiene que ver con la función de las condiciones ambientales en el desarrollo de las diferencias individuales de comportamiento, un problema que puede estudiarse mejor dentro de la amplia gama de variación ambiental provisto por culturas muy disímiles. Esta clase de investigación exige instrumentos que puedan aplicarse en condiciones al menos moderadamente comparables en culturas distintas; es necesario buscar salvaguardas contra las interpretaciones erróneas de los resultados obtenidos con esos instrumentos mediante diseños experimentales adecuados y la familiaridad de los investigadores con las culturas o subculturas estudiadas. Para ello se requiere identificar en cualquier cultura determinada las variables experienciales particulares que se puedan relacionar con diferencias sociales significativas en el desarrollo conductual que caracteriza a dicha cultura (J. W. Berry, 1983; Brislin, 1993; Segall, 1983; Whiting, 1976). Un ejemplo destacado, en el campo de la evaluación de la personalidad, de cómo realizar esos esfuerzos de investigación se encuentra en la publicación de las *Cross-Cultural Anxiety Series* que aborda en exclusiva la medición de la ansiedad en culturas diferentes y que han resultado extraordinariamente fructíferas en la expansión de la base de conocimiento sobre este constructo y la forma en que gente de todo el mundo experimenta la ansiedad (véase por ejemplo, Spielberger y Díaz-Guerrero, 1990).

Otro planteamiento importante consiste en formular una prueba en el contexto de una cultura y aplicarla a personas con antecedentes culturales distintos, en cuyo caso debemos evitar el error de considerar a la prueba desarrollada dentro de una singular estructura cultural como un criterio universal para medir la "inteligencia" o cualquier otro constructo. Tampoco debemos asumir que una baja calificación en dicha prueba tiene la misma explicación causal para un miembro de la cultura en la que fue elaborada que para uno de una cultura distinta. Lo que dicha aproximación *nos permite* averiguar es la distancia cultural entre grupos, así como el grado individual de aculturación y su disposición para beneficiarse de las actividades educativas y vocacionales peculiares de la cultura. Algunos investigadores han exagerado el hecho de que el medio cultural en el que crece un individuo afecta las habilidades cognoscitivas y los conocimientos que adquiere. Algunos ejemplos iniciales incluyen

una prueba de reconocimiento de huellas estandarizada en aborígenes australianos (Porteus, 1931) y una Prueba de Dibujo de un Caballo estandarizada en niños de Pueblo Indio (DuBois, 1939). En ambos casos, el grupo cultural en el que se preparó el instrumento destacaba en las comparaciones con otros grupos.

La tercera aproximación consiste en el desarrollo en cada cultura de pruebas distintas (o adaptaciones sustanciales de las existentes), que se validan contra criterios locales y se emplean únicamente en la cultura apropiada. Esta aproximación la ilustra la preparación de pruebas para personal industrial y militar de culturas particulares. Un ejemplo es el programa de formulación de pruebas conducido en varias naciones en desarrollo de África, Asia y Latinoamérica con el patrocinio de la Oficina Estadounidense para el Desarrollo Internacional (*United States Agency for International Development*) (Schwarz y Krug, 1972). En estos casos, las pruebas se validan contra los criterios educativos y vocacionales específicos que pretenden predecir y la ejecución se evalúa en términos de normas locales. Cada prueba se aplica únicamente en la cultura para la que se creó y no se pretende realizar comparaciones transculturales; sin embargo, si los criterios que se pretende predecir son tecnológicos, es probable que se requiera de una "inteligencia de tipo occidental" y que las pruebas reflejen la dirección en que evoluciona la cultura particular más que las características prevalentes en ese momento. Además, una revisión reciente del uso global de las pruebas indica que las que se utilizan con mayor frecuencia, al menos con niños y jóvenes, son las editadas en Estados Unidos y Europa. De hecho, entre todas las categorías de naciones participantes en el estudio, se descubrió que los países que más utilizan la tecnología psicométrica extranjera son los menos desarrollados, es decir, justamente los que más difieren de los Estados Unidos y los países europeos (Hu y Oakland, 1991; Oakland y Hu, 1992).[29]

Se cuenta con amplia información sobre el examen psicológico de minorías culturales en sociedades pluralistas como las de Estados Unidos, Israel y Holanda (véase, por ejemplo, Bleichrodt y Drenth, 1991; Duran, 1989; Figueroa, 1990; Hessel y Hamers, 1993; Samuda, Kong, Cummins, Lewis y Pascual-Leone, 1991; Zeidner, 1988). En este texto revisamos ese material en las secciones en que pueda presentársele con mayor claridad, de modo que en el capítulo 18 el interés se dirige a las preocupaciones y responsabilidades éticas y sociales relacionadas con la aplicación de las pruebas a minorías culturales; en los capítulos 6 y 7 vimos los problemas psicométricos del sesgo de la prueba y de las interacciones entre grupo y reactivo; en este capítulo nos ocupamos de los instrumentos desarrollados para la evaluación transcultural y en el capítulo 12 consideraremos los problemas que surgen al interpretar los resultados de pruebas transculturales y las principales tendencias actuales.

En nuestros días, la evaluación multicultural se aparta de la elaboración de pruebas especiales para concentrarse cada vez más en la función del examinador en el proceso de evaluación, durante el cual tiene la responsabilidad de: (1) obtener información sobre los antecedentes culturales del examinado, (2) elegir la prueba más adecuada para sus propósitos, (3) presentar y aplicar el instrumento de manera eficaz

[29] En reconocimiento de esta realidad, la Comisión Internacional de Tests preparó un conjunto coherente y cuidadosamente elaborado de directrices para adaptar las pruebas educativas y psicológicas (Hambleton, 1994, 1996; Van de Vijver y Hambleton, 1996). Geisinger (1994) aborda muchos de esos temas.

al individuo y (4) interpretar los resultados de la prueba a la luz de los antecedentes experienciales del individuo y del contexto (ocupacional, educativo, comunitario, etc.) en el cual es evaluado. Estudiaremos con más detalle estos rasgos de la función del examinador en el capítulo 12.

La evaluación del ambiente. Aunque la revisión de las pruebas transculturales resulta de interés histórico y mejora la comprensión del origen y la naturaleza de las pruebas actuales, el rápido aumento en el contacto entre las culturas de todo el mundo está modificando la necesidad de dichas pruebas. Cada vez son más las pruebas eficaces que se desarrollan o adaptan en culturas particulares para propósitos específicos, por ejemplo para usos educativos, industriales y de consejería. La búsqueda de un test universal de inteligencia humana se reconoce como fútil porque cada vez es mayor la conciencia de la importante contribución del ambiente del individuo y su historia de experiencias a la forma que adopta su inteligencia. Estas condiciones han tenido como consecuencia un aumento en la actividad de evaluación del ambiente en el que opera el individuo.[30]

La aproximación tradicional a la evaluación del ambiente se fundamenta en un indicador compuesto, global, del nivel socioeconómico. Los sociólogos suelen emplear complejos procedimientos para identificar la pertenencia del individuo a una clase social (Warner, Meeker y Eells, 1949), pero otros índices más sencillos y de más fácil aplicación han demostrado ser igualmente eficaces y producir resultados que concuerdan estrechamente con los que se obtienen con métodos más laboriosos. De hecho, puede encontrarse una aproximación razonable al nivel socioeconómico a partir de la ocupación del principal proveedor de la familia. Se han compuesto algunas escalas aproximadas para clasificar la ocupación de los padres en niveles, algunas combinan la información sobre la ocupación con la del nivel educativo de los padres, como en el Índice Bifactorial de la Posición Social (*Two-Factor Index of Social Position*), descrito primero por Hollingshead (1957) y que puede encontrarse en varias fuentes (por ejemplo, Bonjean, Hill y McLemore, 1967; Hopkins y Stanley, 1981), aunque también se han desarrollado otros procedimientos más objetivos para registrar la información sobre la ocupación y conseguir un índice del nivel ocupacional (Duncan, 1961; Stricker, 1985).

Una limitación importante de los índices globales tradicionales proviene del hecho de que clasifican los ambientes en continuo de mejor a peor o de superior a inferior, cuando en realidad difieren en las conductas particulares que refuerzan y, por ende, en sus efectos sobre características concretas del individuo (véase, por ejemplo, McAndrew, 1993). Esto supone que los entornos óptimos para el desarrollo de habilidades atléticas, aprovechamiento académico, creatividad y conformidad social pueden ser muy diferentes. En la revisión que el trabajo pionero de Roger Barker hace de la psicología ecológica (Schoggen, 1989), es posible encontrar orientaciones valiosas sobre la aproximación empírica a la clasificación y descripción de los medios conductuales.

La evaluación transcultural realza la importante función que el estilo paterno y el ambiente familiar cumplen en el progreso intelectual del niño (véase, por ejemplo,

[30] Veremos este punto más a fondo en el capítulo 12.

M. H. Bornstein, 1991). En la actualidad, se reconoce que las diferencias ambientales no se limitan a poblaciones culturales o étnicas identificables, sino que pueden afectar significativamente el desarrollo psicológico de cualquier persona. Más aún, dada la necesidad de identificar mejor los ambientes en términos de las conductas particulares que fomentan, cada vez se concede más atención a la evaluación precisa de los efectos psicológicos de diferentes hogares y ambientes familiares.

Se dispone ahora de muchas medidas y pruebas para la evaluación de familias y ambientes familiares (Bradley y Brisby, 1993; Paget, 1991). Un conocido instrumento para la evaluación del ambiente familiar es el Inventario de Observación en el Hogar para la Medición del Ambiente (*Home Observation for Measurement of the Environment*, HOME, B. M. Caldwell y Bradley, 1984). Este instrumento se concentra en las formas de estimulación y conducta de los padres en el entorno familiar que favorecen el desarrollo cognoscitivo (Bradley y Caldwell, 1984; B. M. Caldwell y Bradley, 1978; J. H. Stevens y Bakeman, 1985). Existen tres versiones del inventario HOME diseñadas para emplearse del nacimiento a los tres, de los tres a los seis y de los seis a los 10 años. El HOME proporciona calificaciones en varias escalas que evalúan variables como el suministro de materiales adecuados de juego, la variedad de estimulación, la estimulación lingüística, el fomento de la madurez social y la estimulación de la conducta académica (para una revisión, véase Boehm, 1985). En la niñez temprana, los índices de la posición socioeconómica de la familia (*socioeconomic status*, SES) de los infantes muestran una correlación con la capacidad intelectual tan bueno o mejor que las calificaciones del HOME; sin embargo, en algunas circunstancias la combinación de las calificaciones SES y HOME puede mejorar la predicción de la inteligencia (véase, por ejemplo, D. L. Johnson *et al.*, 1993); además, las variables evaluadas por el HOME e instrumentos similares pueden aportar información única y valiosa a la evaluación de los niños para muchos otros propósitos.

Capítulo Diez

Pruebas colectivas

Mientras que las pruebas individuales, como las escalas de Stanford-Binet y las de Wechsler, encontraron su aplicación principal en el área clínica, los instrumentos colectivos se utilizan sobre todo en el sistema educativo, el gobierno, la industria y los servicios militares. Recuerde que la evaluación masiva tuvo su inicio durante la Primera Guerra Mundial con la elaboración de las pruebas Alfa y Beta para uso del ejército estadounidense. La primera era una prueba verbal diseñada con propósitos de identificación general y asignación; la otra era un instrumento sin lenguaje para personas a las que no se podía aplicar la Alfa, fuera porque hablaran un idioma extranjero o porque fuesen analfabetas. La pauta establecida por esos instrumentos continuó en el desarrollo posterior de numerosas técnicas de evaluación colectiva para uso civil.

Durante varias décadas siguieron en uso las revisiones de las formas civiles de las dos pruebas originales del ejército, donde se preparó el Test de Calificación de las Fuerzas Armadas (*Armed Forces Qualification Test*, AFQT) como instrumento de identificación preliminar al que seguían baterías de clasificación de aptitudes múltiples para la asignación a las especialidades ocupacionales. El AFQT proporciona una única puntuación basada en el mismo número de reactivos de vocabulario, aritmética, relaciones espaciales y de capacidad mecánica. Más tarde se elaboró la Batería de Aptitud Vocacional de las Fuerzas Armadas (*Armed Services Vocational Aptitude Battery*, ASVAB) para emplearse como medida compuesta de selección y clasificación, y algunos de sus subtests sirven como el componente común de calificación. Para la clasificación de personal, cada rama de las fuerzas armadas selecciona y combina las subpruebas que se ajusten a sus necesidades específicas de especialidades ocupacionales.

En este capítulo consideraremos primero las principales diferencias entre las técnicas colectivas y las individuales. Después, haremos una revisión de los procedimientos de

reciente aparición que permiten ajustar individualmente las pruebas de grupo y el uso de computadoras en los programas de evaluación. Luego analizaremos algunos ejemplos actuales de los instrumentos colectivos de amplio alcance para uso general. Por último, nos ocuparemos de una importante orientación en el uso y desarrollo de las pruebas que puede discernirse tanto en los instrumentos grupales como en los individuales revisados en el capítulo ocho. Se trata de una tendencia a combinar las técnicas diseñadas originalmente como medidas globales de una capacidad general (como la inteligencia o la aptitud académica) con baterías de aptitudes múltiples. Cada vez son más los instrumentos de habilidad que se adaptan para permitir un uso flexible, gracias a que un instrumento puede proporcionar calificaciones de diversa amplitud (de general a específica) para adaptarse a una diversidad de situaciones y propósitos de evaluación.

LAS PRUEBAS COLECTIVAS EN COMPARACIÓN CON LAS INDIVIDUALES

Diferencias características en el diseño de las pruebas. Las técnicas grupales difieren de las individuales tanto en la forma como en la disposición de los reactivos. Aunque es posible utilizar preguntas que requieran respuestas abiertas (y de hecho así se hizo en las primeras pruebas colectivas), en la actualidad la prueba colectiva clásica emplea *reactivos de opción múltiple*, un cambio desde luego necesario para asegurar la uniformidad y objetividad de la puntuación. Otra diferencia importante entre los instrumentos tradicionales, individuales y colectivos, estriba en el control de la dificultad de los reactivos. En las individuales, el evaluador sigue reglas de inicio, así como reglas basales y de techo, para asegurar que cada individuo es examinado con los reactivos adecuados para su nivel de habilidad. En las pruebas colectivas, se disponen reactivos de contenido similar en orden creciente de dificultad en *subtests que se cronometran por separado*, lo que permite al individuo que intente resolver cada tipo de reactivos (como vocabulario, aritmética y espacial) y completar los más sencillos de cada categoría antes de tratar de resolver los más difíciles, en los que podría desperdiciar mucho tiempo.

Sin embargo, al separar los subtests puede enfrentarse el problema de que los evaluadores menos experimentados o cuidadosos cometan errores de cronometración, sobre todo cuando el instrumento incluye varios límites temporales cortos en lugar de uno solo largo. Algunos instrumentos utilizan el *formato de espiral colectiva* en un intento por conciliar el uso de un único límite de tiempo con una disposición que permita a todos los individuos tratar de resolver todos los reactivos con grados de dificultad crecientes. Uno de los primeros en introducir este formato fue el Test Autoaplicado de Habilidad Mental de Otis (*Otis Self-Administering Tests of Mental Ability*) que, como lo indica el nombre, pretende reducir al mínimo la participación del evaluador. En una prueba con el formato de espiral colectivo, se presentan primero los reactivos más sencillos de cada tipo, seguidos de los que se encuentran en el siguiente grado de dificultad, y así sucesivamente, en una espiral creciente del grado de dificultad, como se ilustra a continuación:

1. Lo opuesto al odio es: Respuesta
   1. Enemistad, 2. Temor, 3. Amor, 4. Amistad, 5. Gozo . . . . ( )

2. Si tres lápices cuestan 25 centavos, ¿cuántos lápices pueden comprarse con 75 centavos? ..... ( )
3. Un pájaro no siempre tiene:
   1. Alas, 2. Ojos, 3. Patas, 4. Un nido, 5. Una billetera .... ( )
4. Lo opuesto del honor es
   1. Gloria, 2. Desgracia, 3. Cobardía, 4. Temor, 5. Derrota ...( )

Algunos instrumentos aplican la disposición en espiral colectiva a bloques de cinco a 10 reactivos para no repetir las instrucciones en cada uno y así reducir los cambios en el conjunto de instrucciones requerido por los examinados.

Ventajas de las pruebas colectivas. Las pruebas colectivas se diseñan principalmente para la evaluación masiva, y en comparación con las individuales tienen tanto ventajas como desventajas. Por el lado positivo, las pruebas colectivas pueden aplicarse a la vez a tantas personas como puedan acomodarse en el espacio disponible y que se encuentren al alcance de un micrófono. Las técnicas de evaluación colectiva utilizan reactivos impresos que requieren respuestas simples que el examinado puede registrar en la forma u hoja de respuestas, lo que eliminó la necesidad de una relación individual entre evaluador y examinado e hizo posible el desarrollo de los programas de evaluación a gran escala.

Las pruebas colectivas también facilitaron la evaluación masiva al simplificar la función del evaluador. La diferencia en la capacitación y experiencia requeridas para aplicar, por ejemplo, el Stanford-Binet, es que casi todos los instrumentos colectivos requieren sólo la habilidad de leer instrucciones sencillas a los examinados y llevar con precisión el tiempo. Por supuesto, es deseable contar con algunas sesiones previas de capacitación, ya que los evaluadores inexpertos pueden desviarse inadvertidamente del procedimiento estandarizado e influir en los resultados de la prueba. Al disminuir la participación del evaluador, la evaluación colectiva proporciona condiciones más uniformes que la individual. Esta fuente de varianza puede disminuirse aún más con el uso de instrucciones grabadas y con la aplicación computarizada, condiciones que permiten estandarizar el procedimiento de evaluación a gran escala. La puntuación suele ser más objetiva en la evaluación colectiva e incluso un empleado puede hacerlo, aunque en la actualidad casi todas las pruebas colectivas puede calificarlas una computadora.

Desde otro ángulo, en general las pruebas colectivas proporcionan normas mejor establecidas que las individuales. Durante el proceso de estandarización se acostumbra examinar grandes muestras representativas gracias a la relativa facilidad y rapidez con que se obtienen los datos en las pruebas colectivas; en las más recientemente estandarizadas, no es raro que las muestras normativas consten de 100 000 a 200 000, en contraste con los 1 000 a 8 000 casos acumulados de manera laboriosa al estandarizar las escalas individuales de inteligencia preparadas con más cuidado.

Desventajas de las pruebas colectivas. Aunque las pruebas colectivas tienen varias características deseables y cumplen una función casi indispensable en la evaluación actual, también deben señalarse sus limitaciones. En la evaluación colectiva el evaluador tiene menor oportunidad de establecer *rapport* y de obtener y mantener la

cooperación y el interés de los examinados. En los instrumentos colectivos resulta más difícil que en los individuales detectar cualquier condición temporal de los examinados, como enfermedad, fatiga, preocupación o ansiedad, que pueda interferir con su ejecución. En general, para quienes no están acostumbrados a las evaluaciones, un examen colectivo puede suponer mayores desventajas que uno individual, y al respecto hay evidencias que indican que los niños con perturbaciones emocionales pueden desempeñarse mejor en las pruebas individuales que en las colectivas (Bower, 1969; Willis, 1970).

Desde otro punto de vista, se ha criticado a los instrumentos de evaluación colectiva por las restricciones que imponen a las respuestas del examinado. Las críticas se dirigen en particular a los reactivos de opción múltiple y otros como las analogías, las semejanzas y los de clasificación (Hoffman, 1962; LaFave, 1966). Algunos de sus argumentos son agudos y provocativos; por ejemplo, se afirma que estos reactivos pueden castigar a un pensador brillante y original que ve implicaciones inusuales en las respuestas. Entre paréntesis, advierta que, en vista de los datos de validez y del análisis de reactivos, éste sería un caso extraño, e incluso si llegara a ocurrir en uno o dos reactivos de un instrumento individual, difícilmente tendría un efecto apreciable sobre la calificación total del examinado. Otras críticas se han concentrado en la importancia de analizar los errores y averiguar las razones por las que un individuo elige una respuesta particular, como es el caso de la aproximación piagetiana (Sigel, 1963). Es indudable que las pruebas colectivas ofrecen poca o ninguna oportunidad de hacer observaciones directas de la conducta del examinado o de identificar las causas de un comportamiento atípico. Entre otras razones, es por eso que cuando deben tomarse decisiones importantes sobre individuos resulta conveniente complementar las pruebas colectivas con el examen individual de los casos dudosos o con información proveniente de otras fuentes.

Otra limitación de la evaluación colectiva tradicional es su falta de flexibilidad. En la medida en que a todos los examinados se les aplican todos los reactivos, el tiempo disponible para la prueba podría utilizarse más efectivamente si cada individuo se concentrara en las partes apropiadas para su nivel de habilidad; con este procedimiento se evitaría, por un lado, el aburrimiento de trabajar con tareas demasiado sencillas, y, por el otro, la frustración y ansiedad al tratar de resolver reactivos que exceden el nivel actual de habilidad del individuo. En general, las pruebas individuales permiten que el evaluador elija los reactivos en función de las respuestas previas del examinado. Esta distinción entre las pruebas individuales y colectivas es especialmente importante cuando pretenden cubrir un amplio rango de habilidad.

EVALUACIÓN ADAPTATIVA Y APLICACIÓN COMPUTARIZADA

Pruebas individualmente ajustadas. Actualmente se exploran diversas técnicas en un esfuerzo por combinar algunas de las ventajas de la evaluación individual y la colectiva. Hasta ahora el interés se ha concentrado en la manera de ajustar la cobertura de los reactivos a las características de las respuestas de los individuos, aproximación que se conoce en la bibliografía como evaluación adaptativa, secuencial, ramificada,

ajustada, individualizada, programada, dinámica o contingente a la respuesta. Aunque es posible diseñar pruebas colectivas de lápiz y papel que incorporen esos procedimientos adaptativos (Cleary, Linn y Rock, 1968; Lord, 1971a), las técnicas se prestan mejor a la aplicación computarizada del instrumento.

Para la adaptación de los instrumentos pueden seguirse diversos procedimientos (DeWitt y Weiss, 1974; Larkin y Weiss, 1974; Weiss, 1974; Weiss y Betz, 1973). Un ejemplo sencillo incluye el examen de dos etapas que se ilustra en la figura 10.1, en el cual todos los examinados presentan una prueba de ruta compuesta por 10 reactivos que cubren un amplio rango de dificultad. Dependiendo de su ejecución en la prueba de ruta, cada individuo es dirigido a una de las tres pruebas de medición compuestas por 20 reactivos con diferentes grados de dificultad, de modo que aunque la prueba completa tenga 70 reactivos, cada examinado presenta únicamente 30. La prueba piramidal presentada en la figura 10.2 ilustra un arreglo diferente en el que todos los examinados inician con un reactivo de dificultad intermedia. Si su respuesta es correcta, el individuo asciende al siguiente reactivo más difícil, pero si es errónea desciende al siguiente reactivo más sencillo. El procedimiento se repite hasta que el individuo haya dado 10 respuestas. La ilustración muestra una prueba de 10 etapas en la que cada examinado presenta 10 reactivos del total de 55 de que consta la prueba. La línea gruesa muestra la ruta seguida por un examinado cuyas respuestas se asientan en la parte superior como + (correcto) o – (error).

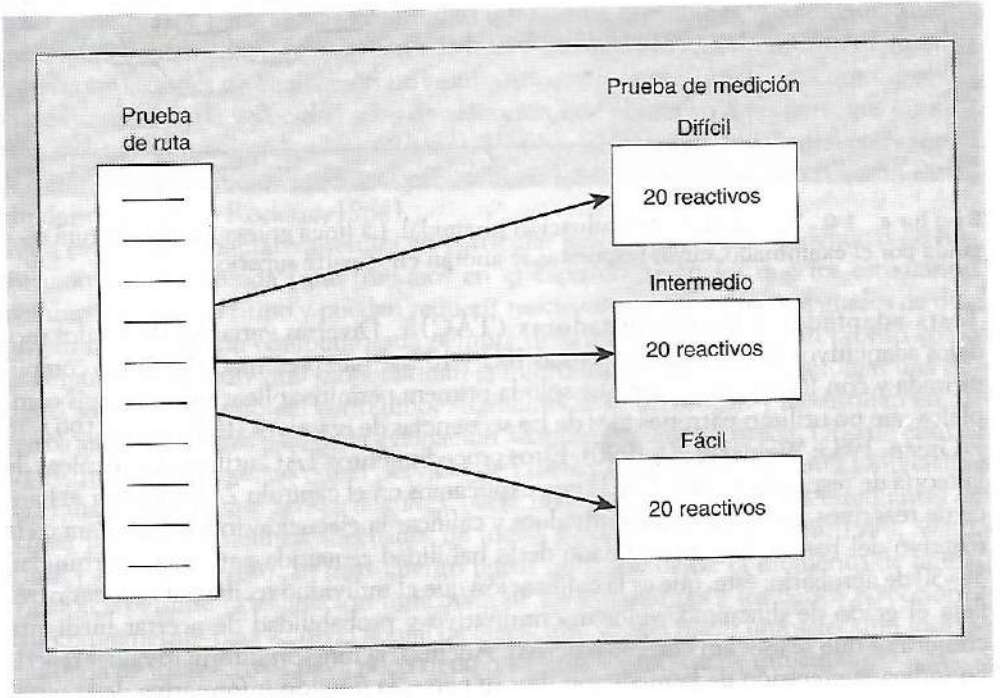


Figura 10.1. Prueba adaptativa de dos etapas con tres niveles de medición. Cada examinado presenta la prueba de ruta y una de las pruebas de medición.

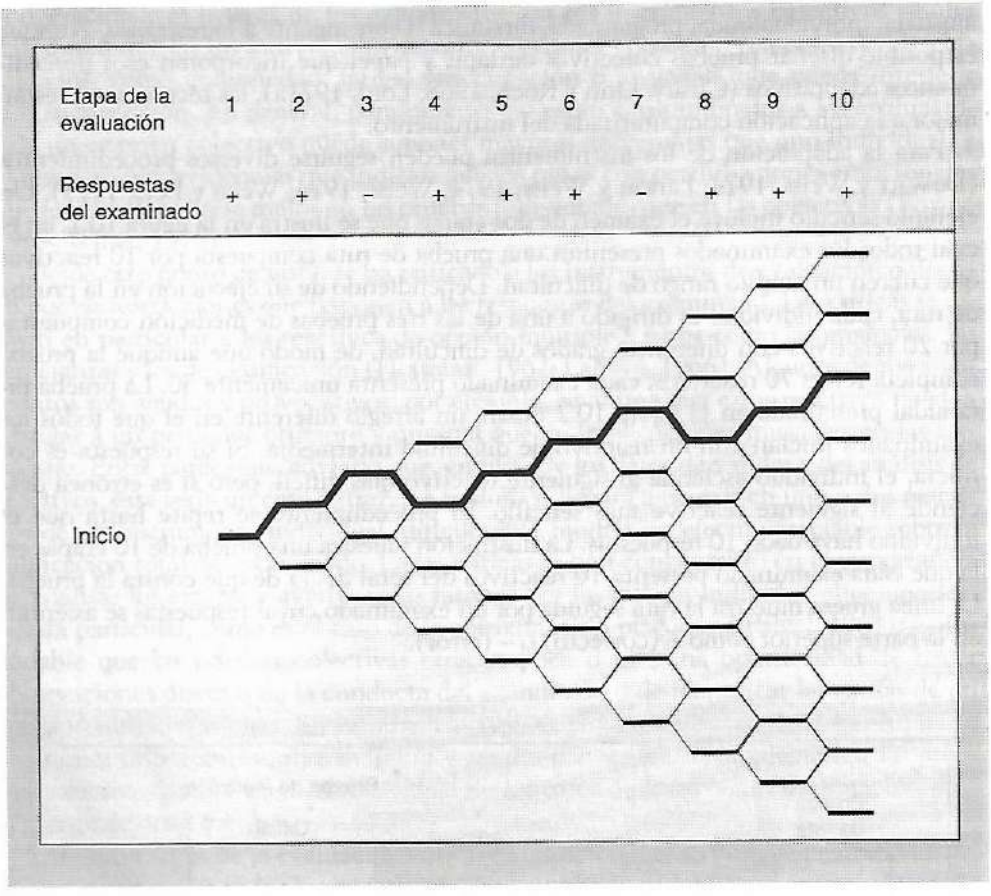


Figura 10.2. Modelo de evaluación piramidal. La línea gruesa muestra la ruta seguida por el examinado, cuyas respuestas se anotan en la parte superior.

Tests adaptados a las computadoras (TAC). Diversas variantes de los dos modelos adaptativos ilustrados en las figuras 10.1 y 10.2 se han presentado en forma computarizada y con lápiz y papel, aunque sólo la primera permite aplicar modelos más complejos que no utilizan patrones fijos de las secuencias de reactivos (Embretson, 1992; B. F. Green, 1983; Wainer *et al.*, 1990). Estos procedimientos TAC utilizan las técnicas de la teoría de respuesta al ítem (TRI) que explicamos en el capítulo 7 para formar el banco de reactivos, examinar a los individuos y calificar la ejecución individual. Para cada reactivo del banco, hay un estimado de la habilidad requerida para una oportunidad 50–50 de aprobarlo; éste, que es la calificación que el individuo recibe por aprobarlo, refleja el grado de dificultad, valor discriminativo y probabilidad de acertar mediante conjeturas que se asocian con ese reactivo. Además, la función informativa del reactivo indica la precisión de la medición. Por su parte, la función informativa de la prueba, que es la suma de las funciones informativas de los reactivos, cumple el mismo propósito que el error estándar de medición (EEM). Después de cada respuesta, la

computadora selecciona el siguiente reactivo de acuerdo con la forma en que el individuo haya respondido hasta ese punto. La evaluación continúa con la adición de nuevos reactivos hasta que la función de información de la prueba alcanza un nivel preestablecido, de modo que todos los individuos son medidos con el mismo grado de precisión.

La calificación individual no se basa en el número de reactivos resueltos correctamente, sino en su grado de dificultad y otras características psicométricas de los reactivos. La calificación total de la prueba se deriva de los estimados de habilidad que corresponden a cada reactivo aprobado. Tales estimaciones se reajustan y perfeccionan cada vez que se agrega un nuevo reactivo hasta alcanzar la precisión predeterminada de la medición. Independientemente del conjunto particular que se aplique a cada individuo, las calificaciones serán equivalentes para todas las personas examinadas con el banco de reactivos. Los procedimientos actuales para la elaboración de un instrumento TAC se facilitan con el uso de diversos programas de cómputo, como MicroCAT, distribuido por la Assessment Systems Corporation.[1]

En general, la investigación realizada con distintos métodos indica que la evaluación adaptativa individualizada puede obtener en menos tiempo la misma confiabilidad y validez con un número mucho menor de reactivos y proporcionar una medición más precisa de los individuos que se encuentran en los extremos superior e inferior del rango de habilidad que cubre la prueba (Lord, 1970, 1971a, 1971b, 1971c; Weiss, 1982). Otras investigaciones indican que la correlación entre una buena versión TAC y su equivalente de lápiz y papel es casi tan elevada como los coeficientes de confiabilidad de la mayor parte de las pruebas, lo que indica que ambas modalidades miden esencialmente los mismos constructos (Mead y Drasgow, 1993). Al mismo tiempo, en algunas situaciones de evaluación no resulta adecuada la versión TAC, como es el caso de las pruebas de velocidad y las de selección que clasifican a las personas en función de una puntuación de corte (Wainer, 1993b). Se ha prestado especial atención al desarrollo de lineamientos técnicos para evaluar los instrumentos CAT (Green, Bock, Humphreys, Linn y Reckase, 1984).

La evaluación adaptativa resulta en particular apropiada para los programas de instrucción individualizada (que tratamos en el capítulo 3) en los que los estudiantes avanzan a su propio ritmo y pueden requerir reactivos con muy diversos grados de dificultad. La evaluación computarizada permite detener la evaluación tan pronto como las respuestas del individuo proporcionan la información suficiente para tomar una decisión sobre el dominio del contenido. Actualmente se investigan y desarrollan en diversos ambientes otros usos de la evaluación adaptativa computarizada. Uno de ellos es un instrumento elaborado por el Servicio de Pruebas Educativas y la Junta Universitaria (*Educational Testing Service and the College Board*) para asignar a los estudiantes de primer ingreso a los cursos adecuados de inglés y matemáticas (Smittle, 1990; Ward, Kline y Flaugher, 1986). El instrumento puede aplicarse desde el momento de la inscripción y permite la asignación inmediata de los estudiantes a los cursos o secciones más convenientes para su nivel actual de ejecución, ya que hacer una selección individualizada de los reactivos lleva poco tiempo y puede calificarse de inmediato.

[1] Las direcciones aparecen en el apéndice B. Véase también Quan, Park, Sandahl y Wolfe (1984) y Weiss y Vale (1987).

El TAC también puede aplicarse en la selección de personal a gran escala y en los programas de clasificación en la industria, el gobierno y los servicios militares. En estos casos resulta especialmente apropiado por tres razones: (1) puede examinarse al flujo estable de solicitantes, lo que impide que se concentren grupos demasiado grandes en cualquier momento y lugar; (2) abarca un amplio rango de niveles de habilidad, y (3) aumenta la seguridad de la prueba porque a cada solicitante se le aplica un conjunto diferente de reactivos de un enorme banco almacenado en la computadora. Desde hace algún tiempo está en progreso una investigación explicatoria orientada al desarrollo de una versión TAC de la Batería de Aptitud Vocacional de las Fuerzas Armadas (*Armed Services Vocational Aptitude Battery*, ASVAB) (McBride y Martin, 1983; Moreno, Wetzel, McBride y Weiss, 1984; Wiskoff y Schratz, 1989). Gradualmente se desarrollan las versiones TAC para las principales pruebas colectivas, como los Tests de Aptitud Diferencial[2] (*Differential Aptitude Tests*) que describiremos en la última sección de este capítulo. El TAC ofrece indudables ventajas para muchas aplicaciones prácticas y permite realizar una investigación permanente sobre la naturaleza y las fuentes de las diferencias individuales. Embretson (1992) hace una presentación clara y efectiva de esta promesa para el futuro de la evaluación.

BATERÍAS DE NIVELES MÚLTIPLES

Generalidades. A diferencia de las principales escalas individuales y de los tests adaptativos computarizados, los instrumentos colectivos *tradicionales* presentan los mismos reactivos a todos los examinados independientemente de sus respuestas, por lo que tienen que cubrir un rango de dificultad reducido y adecuado para la edad, grado o nivel de habilidad para el que se diseñaron. Algunas series actuales de baterías de niveles múltiples traslapados proporcionan medidas comparables del desarrollo intelectual en un rango más amplio, lo que permite, por ejemplo, que cualquier individuo sólo sea examinado con el nivel adecuado para ese momento (los otros niveles pueden usarse al volver a examinarlo años más tarde o para hacer evaluaciones comparativas de diferentes grupos de edad). El hecho de que las baterías sucesivas se traslapen ofrece un piso y un techo adecuados para los individuos que se encuentran en los extremos de sus propias distribuciones por grado o edad. Por supuesto, debe reconocerse que, en el mejor de los casos, las baterías de niveles múltiples permiten una igualación aproximada entre la dificultad del reactivo y la habilidad del examinado. Más aún, a diferencia de los procedimientos individualizados del TAC, la igualación no se basa en las respuestas de los examinados al instrumento, sino en el conocimiento previo sobre ellos, como su edad o grado escolar.

El uso de las baterías de niveles múltiples es particularmente adecuado para las escuelas, ya que ahí se requiere hacer la comparación de las calificaciones a lo largo de varios años. Es por esto que los niveles suelen describirse en términos de calificaciones. La mayor parte de estas baterías brinda un grado razonable de continuidad en relación con el contenido o las funciones intelectuales cubiertas. Las calificaciones se expresan en términos de las mismas unidades de la escala y se utilizan los procedimientos de la teoría de respuesta al ítem (TRI, que estudiamos en el capítulo 7) para permitir la continuidad y la comparación a lo largo del rango. En el proceso de estan-

[2] DAT-Adaptada (1987), correspondiente a la DAT forma V de lápiz y papel (1981).

darización de la prueba, se aplican a grupos de estudiantes niveles traslapados del instrumento para obtener los datos de vinculación que se requieren. Las muestras normativas examinadas a distintos niveles son también más equivalentes de lo que puede ser el caso de las pruebas estandarizadas independientemente. Por lo general, los niveles individuales cubren de uno a tres grados. Una batería de niveles múltiples puede evaluar de manera uniforme un rango que se extiende del jardín de niños al ingreso a la universidad.

La mayor parte de las baterías produce una calificación estándar total que corresponde al tradicional CI de las pruebas individuales. Algunas baterías proporcionan varias normas que incluyen percentiles, estaninas o equivalentes de grado y calificaciones estándares. Además de una calificación global total, la mayor parte también produce por separado calificaciones verbales y cuantitativas, o lingüísticas y no lingüísticas. Este detalle coincide con la observación de que la ejecución de un individuo en los subtests verbales puede ser muy diferente del que muestra en otros subtests, en especial en los niveles superiores.

Los nombres de las baterías también resultan de interés. Se utilizan términos como "inteligencia", "habilidad general", "habilidad mental", "madurez mental", "potencial académico" y "capacidad escolar" para designar básicamente el mismo tipo de instrumento. En el vocabulario de los psicómetras, esos términos son de hecho sinónimos e intercambiables. Es digno de mención el hecho de que en la mayor parte de las baterías recientemente desarrolladas o revisadas se ha reemplazado al término "inteligencia" con nombres más específicos, lo que refleja el reconocimiento de que dicho término había adquirido demasiadas acepciones que podrían llevar a interpretaciones erróneas de los resultados. Las baterías de niveles múltiples evalúan las principales habilidades intelectuales que se consideran requisito para el trabajo escolar. Su función principal es evaluar la preparación del individuo para el aprendizaje escolar en cada etapa del proceso educativo.

Baterías representativas. Las baterías presentadas en la tabla 10.1 ilustran la naturaleza y el alcance de las actuales baterías de niveles múltiples. Las elegimos por la novedad de sus más recientes revisiones, la elevada calidad de los procedimientos con que se desarrollaron y el tamaño y representatividad de sus muestras de estandarización. Otro rasgo notable de esas baterías es que cada una se estandarizó al mismo tiempo con una o dos baterías de niveles múltiples de pruebas de aprovechamiento educativo para los mismos grados, que veremos en el capítulo 17. La aplicación de ambas clases de instrumento a las mismas muestras de estandarización hace posible establecer correspondencias entre los dos grupos de puntuaciones, lo que a su vez permite utilizarlos conjuntamente para hacer una exploración cabal del desarrollo académico del estudiante y de las condiciones que lo afectan.

Para estudiar la confiabilidad y la validez de esas baterías se utilizaron los procedimientos adecuados; por ejemplo, se calculó en los mismos grados escolares la confiabilidad Kuder-Richardson de las puntuaciones totales y de las puntuaciones de dos o tres contenidos separados de las baterías, y se encontraron principalmente en .90. Las correlaciones con el retest son altas, lo que indica una estabilidad satisfactoria; las correlaciones con grados escolares y las pruebas de aprovechamiento indican una buena validez predictiva; además, tanto las correlaciones de las calificaciones parciales como los análisis factoriales revelaron la existencia de un gran factor general en cada batería.

Tabla 10-1

Baterías representativas de niveles múltiples

Batería	Grados	Número de niveles	Normalizada conjuntamente con
Test de Habilidad Escolar de Otis Lennon (Otis-Lennon School Ability test,OLSAT, 7 a. ed.)	K–12	7	Serie de Prueba Standford de Aprovechamiento (9a. ed.)
Test de Habilidades Cognoscitivas (CogAT, forma 5)	K–3 3–12	2 8	Tests de Habilidades Básicas Iowa (grados K-9) Pruebas de Aprovechamiento y Destreza (grados 9–12) Test de Desarrollo Educativo Iowa (grados 9–12)
Test de Habilidades Cognoscitivas (2a. ed., TCS/2)	2–12[a]	6	Pruebas de Aprovechamiento de California (5a. ed.) Pruebas Comprensivas de Habilidades Básicas (4a. ed.)

[a] Existe también una forma separada del *Primary Test of Cognitive Skills* (PTCS) con un conjunto diferente de pruebas apropiadas para los niveles de jardín de niños y primer grado

Contenido habitual de la prueba a niveles diferentes. El menor nivel de edad en el que se ha demostrado que es factible aplicar instrumentos colectivos corresponde al del jardín de niños y el primer grado. Por otro lado, para evaluar a preescolares se requiere que el instrumento sea individual para establecer y mantener el *rapport* y para aplicar los tipos de reactivos orales y de ejecución que sean adecuados para esos niños. No obstante, ya para los cinco o seis años es posible aplicar pruebas impresas a pequeños grupos de no más de 10 o 15 niños, aunque todavía es necesario que el evaluador brinde mucha atención individual a los niños para asegurar que sigan las instrucciones, que den vuelta a las páginas del cuadernillo de manera adecuada y para supervisar otros detalles de procedimiento; en caso de ser necesario pueden aplicarse a grupos mayores con la ayuda de uno o dos asistentes.

En general, las técnicas colectivas para el nivel de primaria cubren el jardín de niños y los dos o tres primeros grados de la educación elemental. En esas pruebas, cada niño recibe un cuadernillo que contiene los dibujos y diagramas que constituyen los reactivos del examen, las instrucciones se dan de palabra y por lo común se acompañan de demostraciones. A menudo se incluyen ejercicios de práctica en los que los niños tratan de resolver uno o dos reactivos de muestra y el evaluador o su asistente revisan las respuestas para asegurarse de que entendieron las instrucciones. Los niños marcan sus respuestas en el cuadernillo correspondiente con una crayola o un lápiz suave; la mayor parte de las pruebas sólo requiere que se marque el dibujo correcto de un conjunto, aunque algunas exigen coordinación motora simple, como al dibujar las líneas que unen dos puntos. Como es evidente, las pruebas para el nivel primario pueden no requerir que el niño lea o escriba.

La mayor parte de las baterías de habilidades múltiples incluye pruebas adecuadas para el *nivel de primaria*. En la figura 10.3 se ilustran cuatro de los 10 tipos de reactivos usados en el nivel A del OLSAT que son adecuados para el jardín de niños. En reconocimiento del rápido crecimiento intelectual que ocurre en las edades más tempranas, la última edición del OLSAT ofrece cuatro niveles separados (A, B, C, D) para jardín de niños, primero, segundo y tercer grados, lo que permite una mayor diferenciación que las ediciones anteriores de la batería y otras baterías múltiples. En el nivel A, el evaluador da oralmente todas las instrucciones y, como se muestra en la figura 10.3, el niño responde llenando el círculo debajo del dibujo que corresponde a la respuesta elegida.

La solución de la prueba se lleva alrededor de 75 minutos. Se aplica en dos sesiones, cada una con un receso de cinco minutos luego de los primeros 15 o 20 minutos. También se cuenta con una Prueba Práctica que contiene reactivos e instrucciones similares que pueden aplicarse el día anterior. En la figura 10.3 se muestran reactivos relativamente sencillos que se usan para familiarizar a los niños con los que encontrarán en la prueba. Las explicaciones presentadas en la figura son una versión muy condensada de las instrucciones orales detalladas y de la clara descripción del contenido que acompañan a cada reactivo. También hay algunas pequeñas diferencias en el formato para facilitar la comprensión y retener la atención de los niños más pequeños; por ejemplo, las páginas y los renglones no sólo se identifican con números, sino también con pequeños dibujos de objetos familiares, como una taza, un zapato o unas tijeras, y cada niño recibe una señal que le permite continuar el trabajo en el renglón adecuado.

Las pruebas para el *nivel de la escuela primaria*, que se extienden a partir del tercer o cuarto grados, tienen mucho en común tanto en contenido como en diseño general. Como a estos niveles se supone cierto alfabetismo funcional, el contenido de las pruebas es predominantemente verbal, aunque la mayor parte también abarca problemas aritméticos u otros contenidos numéricos. Además, algunas baterías incluyen pruebas sin lectura diseñadas para evaluar las mismas habilidades de razonamiento abstracto en niños cuyo idioma materno no es el local, que tienen problemas de lectura u otras desventajas educativas.

En la figura 10.4 se ilustran los tipos de reactivos apropiados para el nivel de la escuela elemental. Son característicos de los empleados en los niveles intermedios del Test de Habilidades Cognoscitivas (*Cognitive Abilities Test*, CogAT), que, como indica la tabla 10.1, incluye dos niveles para grados K de tres y ocho niveles del tercer al duodécimo grado. Cada nivel se imprime en un cuadernillo aparte. Los examinados inician la prueba en niveles diferentes y la terminan con conjuntos de reactivos distintos. El instrumento está diseñado de modo que a casi todos los examinados se les apliquen reactivos que para ellos sean de dificultad intermedia, ya que ahí es más efectiva la discriminación.

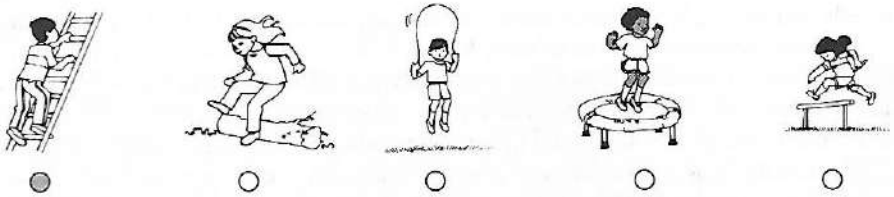
Los niveles A a H contienen los mismos subtests que se agrupan en tres baterías de la siguiente manera:

Batería verbal: Clasificación verbal, completamiento de oraciones, analogías verbales.

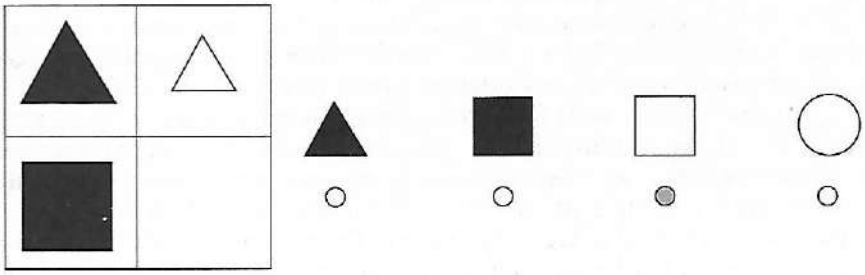
Batería cuantitativa: Relaciones cuantitativas, series de números, desarrollo de ecuaciones.

Batería no verbal: Clasificación de figuras, analogías de figuras, análisis de figuras. Estos subtests no utilizan palabras o números, sino elementos geométricos o figuras, los reactivos tienen poca relación con la instrucción escolar.

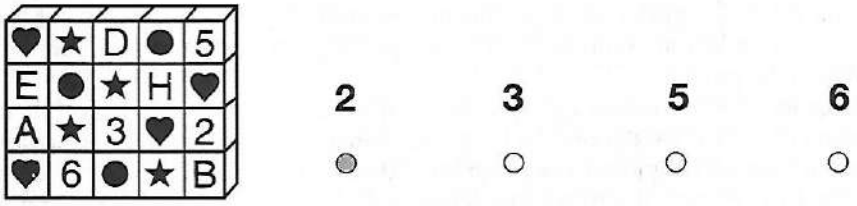
Clasificación de dibujos: Haz una marca en el dibujo que no corresponda con los otros.



Analogía de figuras: Haz una marca debajo de la figura que corresponda al cuadro vacío.



Seguimiento de instrucciones: Marca el número que está justamente debajo de un corazón.



Serie de dibujos: Haz una marca debajo del dibujo que corresponde al cuadro vacío.

Figura 10.3. Algunos de los reactivos de muestra utilizados en el Test de Habilidad Escolar de Otis-Lennon (OLSAT, 7a. ed., Nivel A).

(Copyright © 1996 por The Psychological Corporation. Todos los derechos reservados. Reproducido con autorización.)

Cada subtest está precedido por ejercicios prácticos con explicaciones detalladas, que también están disponibles. La figura 10.4 muestra un reactivo característico de seis de los nueve subtests, con instrucciones condensadas y algo replanteadas, en cuanto al grado de dificultad, que corresponden aproximadamente a los que se aplican de cuarto a sexto grado. El manual recomienda que las tres baterías se apliquen a cada niño en tres sesiones. En la mayor parte de los casos, la batería no verbal no hace una predicción tan buena del aprovechamiento académico como la que pueden hacer las baterías verbal y cuantitativa, pero si se compara la ejecución en las tres puede obtenerse información útil sobre capacidades o discapacidades especiales.

Los niveles superiores de las baterías múltiples, apropiados para *estudiantes de secundaria*, no difieren en esencia de los diseñados para la escuela primaria salvo en el grado de dificultad; también son adecuados para la evaluación general de grupos no seleccionados de adultos con diversos propósitos. El contenido de las pruebas de este nivel puede ilustrarse por los reactivos del nivel más elevado del Test de Habilidades Cognoscitivas (*Test of Cognitive Skills*, TCS/2), que en cada uno de sus niveles contiene cuatro pruebas:

Secuencias: Comprensión y aplicación de una regla o principio a un patrón o secuencia de figuras, letras o números.

Analogías: Identificación de la relación entre un par de imágenes y el completamiento de un segundo par que muestra la misma relación; incluye escenas, gente, animales, objetos o símbolos gráficos.

Razonamiento verbal: Se evalúa con diversos tipos de reactivos, que incluyen la identificación de elementos esenciales de objetos o conceptos, la clasificación de objetos de acuerdo con atributos comunes, la inferencia de relaciones entre conjuntos de palabras o la extracción de conclusiones lógicas de pasajes breves.

Memoria: El sujeto debe estudiar un conjunto de definiciones de palabras artificiales (sílabas sin sentido) y se prueba su recuerdo después de un periodo aproximado de 25 minutos, durante el cual se aplican otros instrumentos.

También se tiene acceso a una prueba de práctica que puede aplicarse uno o dos días antes de la verdadera sesión de evaluación. En la figura 10.5 se presentan ejemplos de tres de cuatro tipos de reactivos. En esta batería se utilizan del cuarto al duodécimo grado los mismos *tipos de reactivos* de las pruebas de secuencias, analogías y razonamiento verbal, y en todos esos niveles se incluyen los mismos reactivos de muestra. A los dos niveles superiores, que corresponden a la secundaria, los distingue la mayor dificultad, empíricamente establecida, de sus reactivos.

La inclusión de una prueba de memoria es un rasgo característico de la batería TCS/2, que proporciona calificaciones separadas para las pruebas verbales, no verbales y de memoria. Esas áreas de aptitud fueron identificadas por los análisis factoriales preliminares que guiaron el desarrollo y la selección de los reactivos de la prueba. El escalamiento entre niveles se obtuvo durante el proceso de estandarización mediante las técnicas de la teoría de respuesta al ítem (TRI, véase el capítulo 7). Con este propósito se aplicaron a muestras de estudiantes pruebas de enlace que contenían reactivos de dos niveles adyacentes (*TCS/2 Technical Report*, 1993, pp. 113–114). La batería TCS/2 hace un uso inusualmente bueno de los procedimientos de TRI tanto en el desarrollo

Clasificación verbal: Piensa en qué se parecen las palabras en negritas y luego encuentra en la línea inferior la palabra que va con ellas.

Amable **amistoso** **servicial**

A capaz B activo C generoso D bonito E fuerte

Analogías verbales: Averigua cómo armonizan las dos primeras palabras y luego encuentra la palabra que completa al segundo par de la misma manera

barco → **puerto : camión** →

A conducir B carretera C cochera D gasolina E carga

Series de números: Averigua la regla utilizada para ordenar la serie de números que se presenta abajo y luego elige el número que sigue.

3 2 1 3 2 1 →

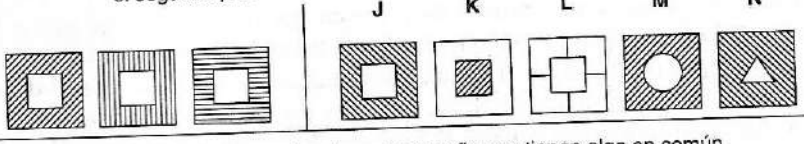
A 0 B 1 C 2 D 3 E 4

Formación de ecuaciones: Los números y signos de la parte superior pueden disponerse de forma que den una de las respuestas de abajo. Marca la respuesta.

2 4 8 - -

J 0 K 2 L 4 M 6 N 10

Analogía de figuras: Decide cómo se relacionan las dos primeras figuras y luego encuentra la figura de la derecha que complete de la misma manera el segundo par.



Clasificación de figuras: Las tres primeras figuras tienen algo en común, encuentra la figura de la derecha que sea como las tres primeras.

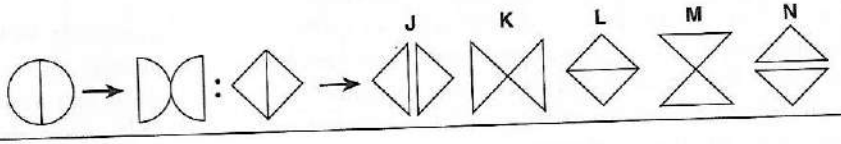
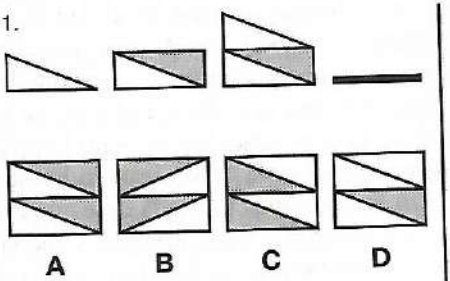


Figura 10.4. Algunos reactivos del Test de Habilidades Cognoscitivas. Las respuestas se marcan en una hoja de respuestas aparte. Las respuestas correctas son: C, C, D, K, J, K.

(Tomado de CogAT, forma 5, Practice Test for Levels A-H. Copyright © 1993 por The Riverside Publishing Company. Reproducido con autorización.)

Secuencias.
Averigua el patrón de cada serie y luego elige el reactivo del renglón inferior que corresponda al espacio en blanco.

1.



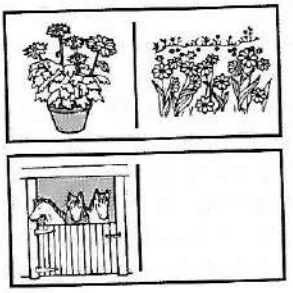
2.

A C E G I ___

H J K L

A B C D

Analogías.
Decide cómo se relacionan los dibujos de los cuadros superiores y luego localiza el dibujo a la derecha que completa de la misma manera el segundo par.



Razonamiento verbal.

1. Mira la palabra subrayada, alfabeto. ¿Cuál de las palabras de abajo nombran algo que siempre forma parte de ella?

   alfabeto

   A palabras
   B letras
   C números
   D enunciados

2. Dada la información de los dos enunciados superiores, ¿cuál de los enunciados de abajo debe ser cierto?

   El Big Ben es un reloj que está en Inglaterra.
   Judy ha visitado el Big Ben.

   A La gente visita a menudo el Big Ben.
   B En Inglaterra hay muchos relojes grandes.
   C El Big Ben recibió su nombre en honor de una persona.
   D Judy ha estado en Inglaterra.

Figura 10.5. Algunos de los reactivos empleados en el Test de Habilidades Cognoscitivas. Las respuestas se marcan en una hoja de respuestas aparte.

(Tomado de TCS/2 Practice Test, Levels 2-6. Copyright © 1992 por CBT/McGraw Hill School Publishing Company. Reproducido con autorización.)

como en la calificación de la prueba, lo que vuelve posible obtener resultados que no sólo reflejen el número de reactivos aprobados, sino también su grado de dificultad.

Reconocimiento de aptitudes múltiples. Como mencionamos al inicio del capítulo, existe una clara tendencia a relacionar la evaluación de una habilidad global general con la medición de aptitudes relativamente independientes. Este acercamiento proviene de lo que inicialmente fueron aproximaciones polémicas e irreconciliables a la evaluación de la habilidad y es análogo a lo que sucedió en el caso de las pruebas individuales, según vimos en el capítulo 8. En este caso, las primeras baterías de niveles múltiples se diseñaron como una versión colectiva de los tests individuales de inteligencia, aunque por lo general tenían la meta más limitada de evaluar la aptitud escolar o la preparación para enfrentar niveles educativos sucesivos. Gradualmente fue evidente que resultaría fructífero complementar una calificación global con otras más limitadas, lo que ocurrió en las tres baterías consideradas en esta sección.

En el OLSAT (edición de 1996) se indicaba que la propia calificación total se limita a un grupo de aptitudes "verbales-educativas" y que la batería no intenta evaluar el segmento "práctico-mecánico" de la inteligencia general. Además, se establecían condiciones para identificar calificaciones complementarias más reducidas de las correspondientes a las partes verbal y no verbal; sin embargo, esta distinción se concentra en los reactivos de ejecución que no requieren el uso del lenguaje en las respuestas, por lo que se dirige sobre todo a la aplicación de instrumentos a estudiantes que tienen un conocimiento limitado del inglés. De cualquier modo, también es posible obtener calificaciones estaninas (por grado o edad) para cinco grupos de pruebas (en las categorías verbal y no verbal) que incluyen comprensión verbal, razonamiento verbal, pictórico, de figuras y cuantitativo. El manual indica que la comparación de la ejecución relativa de un individuo en esos grupos puede ayudar a identificar sus fortalezas y debilidades (*OLSAT, Technical Manual*, 7a. ed., 1997).

El CogAT (forma 5, 1993) no sólo ofrece normas para su calificación total compuesta, sino también para las calificaciones de razonamiento verbal, cuantitativo y "no verbal" (es decir, espacial). Más aún, los informes de las calificaciones individuales incluyen gráficas de barras para esas tres puntuaciones de área y para la calificación total. El manual llama específicamente la atención a la utilidad del perfil de la gráfica de barras para predecir el aprovechamiento académico. Las instrucciones para la interpretación de la calificación hacen reiterada referencia a la importancia práctica de considerar el perfil de calificación del individuo (véase, por ejemplo, *Riverside 2000*, 1994, p. 44), con lo que el reconocimiento de aptitudes múltiples avanza otro paso.

El Test de Habilidades Cognoscitivas, THC/2 (*Test of Cognitive Skills*, TCS/2) es incluso más explícito al reconocer la utilidad de los datos de aptitudes múltiples para la evaluación del rendimiento académico. No sólo se dispone de normas para las calificaciones totales de la batería, sino también para los cuatro subtests y para un total "no verbal" (subtests de secuencias y analogías). Más aún, la propia batería se desarrolló para evaluar tres atributos cognoscitivos amplios identificados mediante análisis factorial: razonamiento verbal, razonamiento no verbal y memoria. Existe entonces un reconocimiento específico de la necesidad de la evaluación de aptitudes múltiples, que veremos en la próxima sección.

MEDICIÓN DE APTITUDES MÚLTIPLES

Los tests tradicionales de inteligencia, individuales o colectivos, se diseñaron con el objetivo principal de obtener una medida global del nivel general de desarrollo cognoscitivo del individuo, como el CI. Sin embargo, los avances teóricos y prácticos pronto dirigieron la atención a la posibilidad de diferenciar ciertas aptitudes en medio del desordenado conglomerado que representaban los primeros tests de inteligencia. Estos avances dieron lugar a la elaboración de instrumentos separados para la medición de unas cuantas aptitudes de gran aplicación y a una definición refinada, así como a una cabal comprensión de lo que miden los tests de inteligencia.

Diversos acontecimientos contribuyeron al aumento del interés por la medición de aptitudes diferentes. En primer lugar, cada vez fue mayor el reconocimiento de la variación intraindividual de la ejecución en los tests de inteligencia. Algunos intentos preliminares por comparar la posición relativa del individuo en los distintos subtests o grupos de reactivos anticiparon en muchos años el desarrollo de las baterías de aptitudes múltiples, pero los instrumentos de inteligencia no se diseñaron con ese propósito. Con frecuencia, la poca confiabilidad de subtests o grupos de reactivos impedía justificar las comparaciones en el mismo individuo. Además, al elaborar los tests de inteligencia por lo común se eligen los reactivos o subtests que permitan obtener una medida unitaria e internamente consistente, es decir, no se pretende incrementar la variación del individuo sino disminuirla al mínimo. En general, se excluyen los subtests o reactivos que tienen una correlación muy baja con el resto del instrumento, precisamente los mismos que deberían conservarse si el propósito fuera diferenciar las habilidades. Por la forma en que se elaboraron casi todos los tests de inteligencia, es poco probable que la ejecución en las mismas pueda ser diferenciada de manera significativa en más de dos categorías, como verbal y no verbal o lingüística y cuantitativa.

El desarrollo de las baterías de aptitudes múltiples fue estimulado por el reconocimiento de que los llamados tests de inteligencia general son menos generales de lo que se suponía al principio. Pronto resultó evidente que muchas de ellas eran más que nada medidas de comprensión verbal y que era común que no abarcaran áreas como la de las habilidades mecánicas, excepto en algunas de las escalas de ejecución y sin lenguaje. Los psicólogos empezaron a calificar el término "inteligencia" cuando se hicieron manifiestas estas limitaciones de los tests de inteligencia, y algunos sugirieron distinciones entre inteligencia "académica" y "práctica", mientras que otros hablaban de inteligencia "abstracta", "mecánica" y "social". También se crearon instrumentos de medición de "aptitudes especiales" para complementar los tests de inteligencia. Pero los análisis más detallados demostraron que los propios tests de inteligencia miden cierta combinación de aptitudes especiales, como la verbal y la numérica.

El desarrollo de las actividades de los psicólogos en la consejería vocacional y en la selección y clasificación del personal industrial y militar dio un fuerte impulso a la evaluación de aptitudes diferenciales. Un reflejo de este interés es el desarrollo de pruebas especializadas en el trabajo de oficina, mecánico y otras áreas vocacionales. Una labor similar que ha estado vigente por muchos años se encuentra en la preparación de baterías para seleccionar a los aspirantes a las escuelas de medicina, leyes, ingeniería, odontología y otros campos profesionales. Además, un resultado directo de la elección

ocupacional o clasificación de trabajo fue la elaboración de baterías de aptitudes diferenciales como las preparadas por las fuerzas armadas y el Servicio Estadounidense de Empleo.

Por último, el estudio de la organización de rasgos mediante las técnicas de *análisis factorial*[3] estableció la base teórica para la elaboración de baterías de aptitudes múltiples. Gracias a dicha investigación sería posible identificar, clasificar y definir más sistemáticamente las distintas habilidades agrupadas en desorden bajo el término "inteligencia". Esto haría posible seleccionar los instrumentos de tal forma que cada uno represente la mejor medida disponible de uno de los factores o rasgos identificados por el análisis factorial.

Tests de aptitud diferencial. Una de las baterías de aptitudes múltiples de mayor uso son los Tests de Aptitud Diferencial (*Differential Aptitude Tests*, DAT), que desde su publicación en 1947 ha pasado por revisiones periódicas (5a. ed., forma C, 1992). La batería se diseñó principalmente para emplearse en la consejería educativa y profesional de estudiantes de octavo a duodécimo grados. Existen dos niveles de la quinta edición del DAT: el nivel uno, diseñado sobre todo para estudiantes de séptimo a noveno grados y para adultos con ese grado de escolaridad, y el nivel dos, para estudiantes de décimo a duodécimo y para adultos con más de nueve años de escolaridad pero que no se graduaron de la secundaria.

Las ocho pruebas que componen el DAT son: Razonamiento Verbal, Razonamiento Numérico, Razonamiento Abstracto, Velocidad y Exactitud Perceptual, Razonamiento Mecánico, Relaciones Espaciales, Ortografía y Uso del Lenguaje. En la figura 10.6 se muestran reactivos de cuatro de esas pruebas. Además, se dispone de una Prueba de Práctica que cubre las ocho áreas y que puede aplicarse previamente a todos los examinados, así como de una forma especial, los Tests de Aptitud Diferencial para la Evaluación de Personal y de Carrera (*Differential Aptitude Tests for Personnel and Career Assessment*), en el que se condensaron e imprimieron los ocho instrumentos en folletos separados, lo que hizo posible la elección y el uso de instrumentos particulares de acuerdo con su pertinencia para ocupaciones específicas.[4]

Igual que casi todas las pruebas actuales, el DAT puede aplicarse en una versión computarizada. Actualmente se explora una forma adaptada del TAC (DAT adaptado) de la que se dispone desde 1987 y en la que, como en toda evaluación adaptada, sólo se aplican al examinado los reactivos apropiados para su nivel de ejecución o capacidad. Para esta versión TAC se usaron los reactivos de una versión anterior del DAT (forma V), los cuales fueron analizados mediante el modelo de Rasch, un modelo simplificado de un parámetro de la teoría de respuesta al ítem (véase el capítulo 7).

A lo largo de los años se ha acumulado una extensa colección de datos sobre la validez del DAT, tanto por parte del editor como de investigadores independientes que han aplicado la batería en distintos ambientes ocupacionales o que la han incluido en proyectos de investigación. La mayor parte de los datos atañe a la validez predictiva

[3] Analizaremos el tema en el capítulo 11.

[4] Un inventario de Intereses Profesionales se estandarizó junto con la quinta edición del DAT. Puede aplicarse, calificarse e interpretarse en combinación con el DAT para la consejería educativa y vocacional.

Razonamiento verbal

Elige el par correcto de palabras para llenar los espacios vacíos. La primera palabra del par va en el espacio en blanco al inicio del enunciado; la segunda va al final del enunciado. .....es a la aleta como el pájaro es a .....

A agua — pluma
B tiburón — nido
C pez — ala
D nadar — volar
E pez — cielo

La respuesta correcta es C

Razonamiento numérico

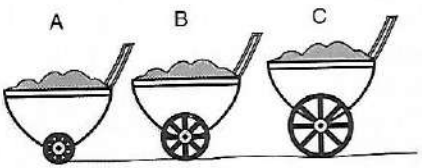
¿Qué número debe reemplazar a R en este problema de suma?

7R	F	G	H	J	K
+R					
88	9	6	4	3	Ninguno

La respuesta correcta es F

Razonamiento mecánico

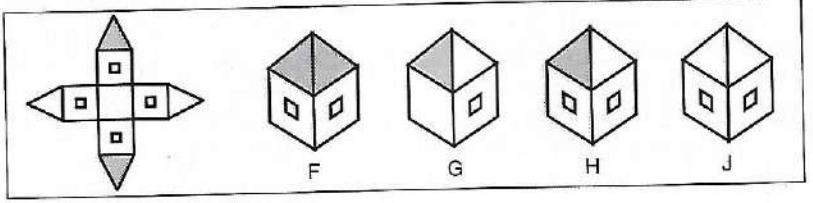
¿Qué carga debe ser más fácil de llevar por arena suave?



La respuesta correcta es C

Relaciones espaciales

¿Cuál de las siguientes figuras podría hacerse doblando el patrón mostrado a la izquierda?



La respuesta correcta es H

Figura 10.6. Reactivos de muestra del Test de Aptitud Diferencial, quinta edición.

(Copyright © 1990 por The Psychological Corporation. Todos los derechos reservados. Reproducido con autorización.)

en términos del aprovechamiento mostrado en programas académicos y vocacionales de la secundaria. Muchos de los coeficientes son elevados, aun con intervalos de hasta tres años entre la prueba y la obtención de los datos del criterio. En lo que respecta a la predicción diferencial los resultados son menos prometedores, pues aunque en general las pruebas verbales tienen una correlación más alta con los cursos de inglés y las numéricas con los cursos de matemáticas, hay evidencias de un gran factor general que fundamenta el desempeño en todo el trabajo académico. Un ejemplo es la elevada correlación que el Razonamiento Verbal muestra con la mayor parte de los cursos, que fue el principal motivo de que se introdujera la puntuación RV + RN como índice de la aptitud escolar, ya que al sumar las puntuaciones obtenidas en los subtests de Razonamiento Verbal y Numérico, el índice muestra una correlación entre .70 y .80 con el criterio compuesto del rendimiento académico. Además de contar con las normas para este índice, que es una de las calificaciones que regularmente se incluyen en el perfil del DAT (véase la figura 4.6), el usuario también dispone de una *Batería Parcial* que sólo contiene los subtests de RV y RN que puede utilizar cuando únicamente desea un indicador general de la aptitud académica.

Desde otro punto de vista, cada vez son más las evidencias de que los tests tradicionales de "inteligencia general" o de "aptitud escolar", colectivas o individuales, proporcionan coeficientes de validez sustanciales en comparación con una amplia variedad de criterios educativos y ocupacionales (L. S. Gottfredson, 1986a; Guion y Gibson, 1988; Pearlman *et al.*, 1980; Schmidt, Hunter, Pearlman y Shane, 1979), es decir, estos instrumentos incluyen en esencia el mismo grupo de habilidades cognoscitivas y conocimientos evaluados por la puntuación RV + RN del DAT. Resulta entonces evidente que la diferencia inicial entre los tests de inteligencia y los de aptitudes múltiples está siendo subsanada desde ambos extremos: por un lado, instrumentos como el DAT dan mayor énfasis al uso e interpretación de puntuaciones de gran amplitud como el RV + RN, mientras que al mismo tiempo que las pruebas inicialmente generales subrayan el uso e interpretación de calificaciones de subtests separados y el análisis de perfiles. Para revisiones del DAT, véase Hattrup (1995) y N. Schmitt (1995).

Batería de aptitudes multidimensionales. La Batería de Aptitudes Multidimensionales (*Multidimensional Aptitude Battery*, MAB) ejemplifica una aproximación aún más cercana al nuevo patrón en la evaluación de las habilidades. Publicada en 1984, sus procedimientos de aplicación, normas y manual han sido actualizados (Jackson, 1994b). La MAB es una prueba colectiva diseñada para evaluar las mismas aptitudes que la Escala de Inteligencia para Adultos de Wechsler, Revisada (WAIS-R, véase el capítulo ocho). Incluye cinco subtests en la Escala Verbal[5] y cinco en la Escala de Ejecución, y proporciona CI de desviación V, E y de la escala completa. La MAB es adecuada para adolescentes y adultos, pero no se recomienda para personas con retardo mental o trastornos mentales cuya condición puede interferir con su comprensión o seguimiento de las instrucciones.

[5] El subtest de retención de dígitos del WAIS-R no tiene equivalente en la MAB. Era difícil presentarlo en un formato de lápiz y papel y su correlación con las calificaciones de la escala completa de Wechsler tendía a ser la más baja.

A continuación se presenta la lista de los 10 subtests, que tienen los mismos nombres que los subtests correspondientes del WAIS-R (con una excepción):

VERBAL	EJECUCIÓN
Información	Símbolos en dígitos
Comprensión	Completamiento de figuras
Aritmética	Espacial
Semejanza	Ordenamiento de figuras
Vocabulario	Ensamble de objetos

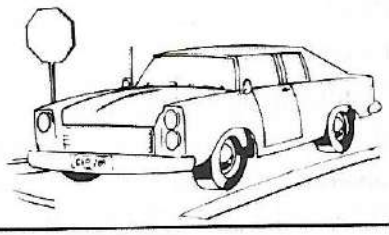
El subtest de Diseño con Cubos del WAIS-R se reemplazó por el subtest Espacial en la MAB. Al diseñar reactivos de lápiz y papel para medir las mismas funciones cubiertas por el WAIS-R se mostró un gran ingenio, en particular en los subtests de ejecución. La figura 10.7 muestra un sencillo reactivo de demostración del subtest de Completamiento de Figuras y uno del subtest Espacial. La tarea del examinado en todos los reactivos de cada uno de los subtests es la misma que la que se ilustra en esos reactivos. En el de Completamiento de Figuras el individuo debe pensar en el nombre de la parte faltante y luego identificar la inicial del nombre entre las opciones dadas. En el subtests Espacial, sólo una de las cinco figuras de la derecha puede obtenerse al girar la figura de la izquierda en una posición diferente sobre la página, mientras que las otras opciones requieren que la figura se voltee.

Los cinco subtests verbales y los cinco no verbales de "Ejecución" se presentan en cuadernillos aparte, cada uno de los cuales empieza con problemas de práctica que ilustran los reactivos empleados en tres de los cinco subtests. Por su parte, cada subtests empieza con uno, dos o tres reactivos más de demostración. En el manual se presentan las instrucciones generales y específicas para cada subtest, que pueden darse oralmente, por medio de una grabación o de una computadora. Las respuestas se registran en hojas separadas o en una computadora, que también puede calificarse en el caso de que no se haga a partir de las plantillas en forma manual. El usuario dispone, entre otras opciones, de una versión de la MAB que puede aplicarse y calificarse completamente por computadora.

Al hacer referencia a la tabla normativa es posible transformar las puntuaciones crudas (o directas) de cada uno de los 10 subtests en calificaciones escaladas uniformes (calificaciones estándares con $M = 50$ y $DE = 10$). Las sumas de esas calificaciones escaladas en V, E y la Escala Completa se utilizan para buscar los CI de desviación ($M = 100$, $DE = 15$) en cada uno de nueve grupos de edad que van de 16 a 74 años. También hay tablas para encontrar en cada uno de los grupos de edad las calificaciones escaladas equivalentes que pueden emplearse para graficar perfiles de edad apropiada. Una ventaja del uso de las calificaciones de una sola batería en lugar de un conjunto de pruebas diferentes es que las de una batería se norman con la misma prueba de estandarización y por ende son directamente comparables.

En conjunto, los procedimientos psicométricos seguidos en el desarrollo de la MAB fueron de alta calidad técnica y cada paso se sustentó en una investigación in-

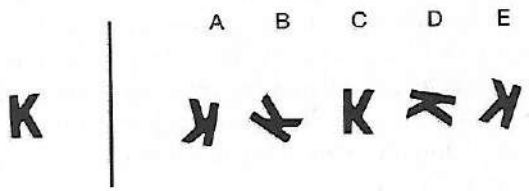
Completamiento de figuras – Elija la letra con la que empieza la palabra que describe a la parte faltante de la figura.



A. L
B. E
C. B
D. W
E. F

La respuesta es **Luz**, por lo que debe marcarse la **A**.

Espacial – Elija una figura a la derecha de la línea vertical que sea igual a la figura de la izquierda. Una de las figuras puede girarse para que se vea igual a la de la izquierda, las otras tendrían que ser volteadas.



La respuesta correcta es **A**, por lo que hay que marcarla.

Figura 10.7. Reactivos de demostración de dos pruebas de ejecución de la Batería de Aptitud Multidimensional (*Multidimensional Aptitude Battery*, MAB).

(Copyright © 1983 por Douglas N. Jackson. Reproducido con autorización.)

tensa que abarcó más de una década.[6] Un rasgo notable de la MAB es su vinculación empírica con el WAIS-R. Primero se demostró que cada subtest de la MAB, así como las calificaciones totales V, E y de Escala Completa, correlacionaban muy alto con las calificaciones del WAIS-R correspondientes en una muestra heterogénea de 145 adolescentes y adultos. De hecho, salvo dos excepciones (Símbolos en Dígitos y Espacial-/Diseño con Cubos), fueron tan altas (o más) como las correlaciones entre WAIS y WAIS-R. Debe advertirse que las correlaciones se obtuvieron a pesar de que los reactivos de la MAB son nuevos y que no existen reactivos comunes entre ambas escalas. Posteriormente se llevó a cabo la igualación lineal de las calificaciones de la MAB con las del WAIS-R en una muestra heterogénea igualada de 160 individuos de 16 a 35 años a los que se les había aplicado las dos pruebas. La igualación preliminar

[6] Véanse las revisiones de S. B. Reynolds (1989) y de Silverstein (1989).

en muestras de estudiantes y de pacientes psiquiátricos indicaba que dicha calibración podía generalizarse a diferentes muestras de examinados. Esos procedimientos de calibración ofrecen una forma prometedora de expresar las calificaciones de las pruebas recién elaboradas en términos de una escala uniforme basada en los datos disponibles de una muestra de estandarización grande y representativa, por lo que son un paso hacia la meta de contar con las normas anclas nacionales de las que hablamos en el capítulo 3.

Desde otro punto de vista, la batería ejemplifica la tendencia actual hacia las calificaciones jerárquicas, puesto que proporciona calificaciones que pueden interpretarse a nivel de los 10 subtests, al nivel más amplio Verbal y de Ejecución o bien al de la calificación global de la batería completa. Así, el usuario tiene la flexibilidad suficiente de elegir el nivel de calificación adecuado a su propósito particular, una condición que probablemente caracterice la evaluación de las habilidades en el siglo XXI.

Capítulo Once

Naturaleza de la inteligencia

El objetivo de todas las pruebas psicológicas es medir la conducta, de ahí que la selección del instrumento apropiado y la interpretación de sus resultados exijan que tanto el que desarrolla el instrumento como el usuario estén familiarizados con la investigación conductual correspondiente. En consecuencia, en este capítulo y el siguiente analizaremos parte de los conocimientos acumulados sobre la conducta que pretenden medir los instrumentos psicométricos con el fin de que el lector entienda la contribución de la investigación psicológica a la comprensión de: (1) la conducta medida por los tests de habilidades cognoscitivas o "inteligencia", (2) las fuentes de variabilidad individual y (3) la predicción del comportamiento en el futuro y en otros contextos.

Para empezar, debe advertirse que el término no calificado de "inteligencia" tiene diversos significados no sólo para el lego, sino también para los estudiosos de diferentes disciplinas como la biología, la filosofía, la educación (véase Sternberg, 1990) e incluso para los psicólogos especializados en diferentes áreas o que se identifican con orientaciones teóricas distintas (por ejemplo, H. Gardner, 1983, 1993; Sternberg, 1985a, 1989; véase también Brody, 1992; Lubinski y Benbow, 1995; Messick, 1992; H. Rowe, 1991). Esta diversidad de significados se hizo evidente ya en 1921, cuando el director del *Journal of Educational Psychology* invitó a 17 de los más destacados investigadores a contribuir con sus definiciones y conceptos de inteligencia ("Inteligencia...", 1921). Luego, 65 años más tarde, se inició un estudio similar (Sternberg y Detterman, 1986). Es obvio que el análisis de esas publicaciones resulta de enorme trascendencia no sólo por la posibilidad de entablar un debate enriquecedor, sino también porque posibilita un punto de convergencia de las opiniones en conflicto.

Aquí, nuestros propósitos son mucho más limitados, pues sólo pretendemos identificar lo que necesitamos conocer acerca del aspecto particular de la inteligencia humana que evalúan los tests tradicionales de inteligencia y que simboliza el CI, por lo que en esta revisión (véase Anastasi, 1983c) emplearemos el término "inteligencia" en el sentido más reducido.

SIGNIFICADO DEL CI

Por lo general la gente no identifica al CI con la calificación de algún instrumento en particular, sino que suele emplear el término como designación abreviada de inteligencia.[1] Esta costumbre se ha vuelto tan común que no podemos limitarnos a ignorarla o deplorarla como una creencia popular errónea. Al considerar el valor numérico de un determinado CI es indispensable especificar de qué prueba se obtuvo, porque los distintos instrumentos que producen tal valor no sólo difieren en su contenido, sino también en otros aspectos que influyen en la interpretación de sus puntuaciones. Aunque en los capítulos anteriores vimos algunos ejemplos de las diferencias entre los instrumentos agrupados bajo el rubro de "tests de inteligencia", es conveniente examinar una vez más las connotaciones generales del constructo de "inteligencia" que simboliza el CI.

Primero, la inteligencia evaluada no debe considerarse como una explicación, sino como una descripción. El CI es sencillamente la expresión del nivel de habilidad que, de acuerdo con las normas de edad disponibles, muestra un individuo en cierto momento, pero no existe un test de inteligencia que pueda explicar las causas de dicho desempeño. Por ende, atribuir el fracaso en un instrumento o en la vida cotidiana a una "inteligencia inadecuada" es una tautología que en nada contribuye a comprender los problemas del individuo y sí, en cambio, nos hace postergar los esfuerzos por explorar en su historia personal las verdaderas causas de su desventaja.

Las pruebas psicológicas en general, y los tests de inteligencia en particular, no deben emplearse como una simple etiqueta para clasificar a los individuos, sino como un instrumento que nos ayude a comprenderlos. Aunque este punto ha sido señalado reiteradamente, lo mismo en trabajos individuales que en los informes de comisiones nacionales (Hobbs, 1975a, 1975b; National Commission..., 1990), un libro de gran difusión (Herrnstein y Murray, 1994), *The Bell Curve*, contribuyó a perpetuar muchos de los estereotipos y las creencias erróneas sobre las diferencias étnicas y de género en la ejecución en los tests de inteligencia, por lo que aumentó la confusión y la controversia sobre un problema ya de por sí complejo. Por su parte, la Fuerza de Tarea sobre la Inteligencia de la Asociación Psicológica Estadounidense (*American Psychological Association Task Force on Intelligence*, Neisser *et al.*, 1966) hizo una aclaración objetiva y real de los puntos importantes, en la que también contribuyó el simposio dedicado al tema durante la convención de la APA en 1995 (Steele, Chair, y August, 1995). Para ayudar a la gente a alcanzar su nivel óptimo de funcionamiento necesitamos comenzar por su punto de partida, evaluar sus fortalezas y debilidades y hacer un plan en función

[1] Inicialmente el CI (cociente de inteligencia) se refería a esta calificación, es decir, la razón de la edad mental y la edad cronológica (véase el capítulo 3).

de ello. Si una prueba de lectura nos indica que cierto niño muestra un retardo en esa función, no podemos limitarnos a etiquetarlo como mal lector o aplicarle una prueba no verbal que disimule la desventaja: debemos concentrarnos en enseñarlo a leer. Más aún, la evaluación contemporánea se ha propuesto la meta de coadyuvar a la comprensión que el individuo tiene de sí mismo y a su desarrollo personal, por lo que cada vez se utiliza más la información de las pruebas para ayudar a la gente a planificar su educación y su carrera y a tomar decisiones personales. Esta aplicación de la evaluación se hace evidente en la atención concedida a las formas efectivas de comunicar los resultados de la evaluación.

El segundo punto que debe recordarse es que la inteligencia no es una capacidad unitaria, sino un compuesto de varias funciones. El término suele emplearse para cubrir la combinación de habilidades que una cultura requiere para sobrevivir y progresar (Anastasi, 1986c). De lo anterior se deduce que las habilidades específicas incluidas en ese compuesto, así como sus pesos relativos, varían según el tiempo y el lugar; por consecuencia, puede esperarse que las condiciones requeridas para un aprovechamiento exitoso difieran en culturas disímiles, en distintos periodos históricos de la misma cultura e incluso en el ciclo vital del individuo, de la infancia a la etapa adulta. Una habilidad relativa tiende a aumentar o a disminuir con la edad en aquellas funciones cuyo valor destaca o menoscaba la cultura o subcultura a la que se pertenezca.

Los tests tradicionales de inteligencia para escolares o adultos miden principalmente las habilidades que predominan en el aprendizaje escolar, como la verbal y, en menor medida, la habilidad para manejar números y otros símbolos abstractos, por lo que puede considerarse que la mayor parte de estos instrumentos mide la aptitud escolar o la inteligencia académica. En ese sentido, el CI es un reflejo del aprovechamiento educativo previo y un predictor del rendimiento educativo subsecuente. Como las funciones enseñadas en el sistema educativo tienen mucha importancia en las culturas modernas tecnológicamente avanzadas, la puntuación de un test de inteligencia académica también predice el desempeño en diversas ocupaciones y otras actividades cotidianas.

Por supuesto, observemos que hay muchas funciones psicológicas importantes que los tests de inteligencia nunca han intentado medir, como las aptitudes mecánica, musical y artística. Las variables motivacionales, emocionales y actitudinales son determinantes importantes del aprovechamiento en todas las áreas; de hecho, algunos psicólogos incluyen componentes de personalidad en su definición de inteligencia (véase H. Gardner, 1983). De igual modo, la investigación sobre la creatividad está identificando variables tanto cognoscitivas como de personalidad que se asocian con la productividad creativa. Desde luego, todo lo anterior implica que tanto las decisiones individuales como las institucionales deben basarse en tantos datos relevantes como razonablemente puedan obtenerse. Basar las decisiones sólo en las pruebas, en especial si son apenas una o dos, equivale sin duda a emplearlas en forma inadecuada. Las decisiones deben ser tomadas por las personas; las pruebas sólo son una fuente de información que deben considerar al decidir, pero no son el instrumento para hacerlo.

Buena parte de la información que poseemos acerca de lo que miden los tests de inteligencia proviene de los estudios sobre la validez de estos instrumentos para predecir el aprovechamiento educativo y vocacional. A un nivel más teórico, el final de la década de los setenta atestiguó el resurgimiento del interés en el constructo de la inteligen-

cia medido por los tests (Humphreys, 1979; Resnick, 1976; Sternberg y Detterman, 1979), interés que no ha perdido vigor y que, por el contrario, abarca diferentes campos de la psicología y es compartido por investigadores con distintas aproximaciones metodológicas y diversas orientaciones teóricas, lo cual se refleja en una serie continua de publicaciones (Detterman, 1985–1993; Sternberg, 1982–1989) así como en una enciclopedia dedicada por completo al tema (*Encyclopedia of Human Intelligence*, 1994).

Los esfuerzos por comprender lo que miden los tests de inteligencia no sólo comprenden los procedimientos estadísticos de validación de constructos, como el análisis factorial, sino también la aplicación de las técnicas de procesamiento de información a las tareas presentadas en dichos instrumentos (véase el capítulo 5). Esta última aproximación no se limita a considerar si la respuesta del sujeto es correcta, sino que pretende identificar los procesos elementales por los que el examinado llega a ella. Es de esperar que este análisis contribuya de manera importante a la utilización de diagnóstico de los instrumentos y a la elaboración de programas de capacitación que permitan satisfacer las necesidades del individuo.

HEREDABILIDAD Y MODIFICABILIDAD[2]

La aplicación de las estimaciones de heredabilidad a las puntuaciones de los tests de inteligencia ocasionó mucha confusión y polémicas. El ejemplo más conocido es un artículo de Jensen (1969) que causó furor y dio lugar a discusiones acaloradas, algunas de las cuales aún continúan. Aunque muchos aspectos de la controversia resultan complicados, una de sus fuentes tiene que ver con la interpretación de las estimaciones de heredabilidad. En concreto, el índice de heredabilidad muestra la contribución proporcional de factores genéticos o hereditarios a la varianza total de un rasgo particular en determinada población en las condiciones actuales; por ejemplo, la afirmación de que la heredabilidad del CI en el Stanford-Binet entre los estadounidenses que asisten a secundarias urbanas es .70 podría significar que 70 por ciento de la varianza encontrada en esas puntuaciones es atribuible a diferencias heredadas y 30 por ciento al ambiente.

Aunque se han empleado diversas fórmulas para calcular los índices de variabilidad (véase, por ejemplo, Jensen, 1969; Loehlin, Lindzey y Spuhler, 1975), sus datos básicos son las mediciones del parecido familiar en el rasgo considerado. Un procedimiento común consiste en utilizar las correlaciones de los tests de inteligencia en gemelos monocigóticos (idénticos) y dicigóticos (fraternos), o bien las correlaciones entre gemelos monocigóticos criados juntos o por separado en hogares adoptivos.

Al interpretar las estimaciones de heredabilidad hay varios puntos que deben tenerse en cuenta. Primero, hay cierta distorsión entre los datos empíricos de las semejanzas familiares debido a que no se evalúan las contribuciones de algunos factores ambientales; por ejemplo, se cuenta con evidencias que indican que el ambiente que comparten los gemelos monocigóticos es más parecido que el que comparten los dici-

[2] La cuestión analizada en esta sección se reduce a una pequeña parte de una extensa área de investigación sobre la operación de la herencia y el ambiente en el desarrollo del comportamiento. El lector puede encontrar una amplia revisión del tema, que incluye la evaluación crítica de los índices de heredabilidad, en los trabajos de Brauth, Hall y Dooling (1991), Bronfenbrenner y Ceci (1994), Horowitz (1994), Plomin y McClearn (1993) y Plomin y Reade (1991).

góticos (Anastasi, 1958, pp. 287–288; Koch, 1966) y que el ambiente de los hermanos criados juntos puede ser psicológicamente distinto (Daniels y Plomin, 1985). Un problema adicional proviene del hecho de que las parejas de gemelos criados por separado no se asignaron al azar a diferentes hogares adoptivos, como sería el caso en un experimento ideal; de hecho, es bien sabido que la asignación a los hogares adoptivos se determina por las características del niño y de la familia que lo adopta, de ahí que sea probable que el ambiente de los hogares adoptivos de los gemelos sea lo bastante parecido para explicar parte de la correlación de las puntuaciones que obtienen en las pruebas. También hay evidencias de que los datos obtenidos con gemelos sobre la heredabilidad pueden no ser generalizables a la población, porque los gemelos son más susceptibles de traumas prenatales que producen retardo mental grave. En consecuencia, si se incluyen en la muestra esos casos de retardo puede aumentar considerablemente la correlación observada de las puntuaciones de los gemelos en los tests de inteligencia (Nichols y Broman, 1974).

A partir de los datos cuestionables, los índices de heredabilidad tienen otras limitaciones intrínsecas (véase Anastasi, 1971; Hebb, 1970) citadas, por cierto, en la primera parte del artículo de Jensen (1969, pp. 33–46). Primera, el concepto de heredabilidad se aplica a poblaciones, no a individuos; por ejemplo, al tratar de establecer la etiología del retardo mental de cierto niño, el índice de heredabilidad no resultaría útil, pues, independientemente del índice en la población, el retardo que este niño muestre puede ser resultado de un gen defectuoso (como en la fenilcetonuria), de daño cerebral prenatal o de privación extrema de estimulación ambiental.

Segunda, los índices de heredabilidad se refieren a la población y al momento en que se encontraron, de ahí que cualquier cambio en las condiciones hereditarias o ambientales pueden alterarlos; por ejemplo, un incremento en la endogamia (como podría suceder en una isla incomunicada) reduciría la varianza atribuible a la herencia y, por ende, el índice de heredabilidad, mientras que, por otro lado, la mayor homogeneidad del ambiente reduciría la varianza atribuible al medio y así elevaría el índice. Además, no puede aplicarse el índice de heredabilidad calculado para una población al análisis de las diferencias en la ejecución en una prueba entre dos poblaciones (como dos diferentes grupos étnicos).

Tercera, la heredabilidad no indica el grado modificable de un rasgo. Incluso si el índice de un rasgo en alguna población es del 100 por ciento, no puede suponerse que la contribución del ambiente es insignificante. Un ejemplo extremo puede ayudar a aclarar este punto. Suponga que una hipotética comunidad de adultos consume la misma dieta, es decir, todos reciben los mismos alimentos en cantidades idénticas. Aquí, la contribución de la comida a la varianza en la salud y la condición física sería nula, ya que la varianza en la alimentación no explica ninguna de las diferencias individuales en estos aspectos; no obstante, si los abastecimientos alimenticios se interrumpieran súbitamente, la comunidad entera moriría de inanición. Por el contrario, mejorar la calidad de la dieta daría por resultado un incremento general en la salud de todos.

Independientemente de la magnitud de los índices de heredabilidad encontrados en los CI de diversas poblaciones, un hecho empírico bien establecido es que el CI no es fijo e inmutable, sino que puede ser modificado por la mediación del ambiente. En el próximo capítulo revisaremos parte de las evidencias a favor de esta conclusión obteni-

das de estudios longitudinales. Por otro lado, se ha avanzado en la identificación de las características de aceleración y desaceleración de los ambientes. El aumento o la disminución en el CI pueden ser resultado lo mismo de cambios ambientales fortuitos en la vida del niño que de intervenciones ambientales planeadas; por ejemplo, el CI puede aumentar o disminuir en forma notable merced a cambios importantes en la estructura familiar, el aumento o la disminución de los ingresos de la familia o la adopción.

Durante las décadas de los setenta y los ochenta se difundió por todo el mundo el interés por la aplicación sistemática de algunos programas para el desarrollo de la inteligencia. Un ejemplo es la edición, de 1980 a 1987, de *Human Intelligence International Newsletter*, publicación que contaba con un consejo editorial internacional y presentaba un informe trimestral de la investigación cognoscitiva y sus aplicaciones a la educación. Otro ejemplo importante fue la creación en Venezuela de un programa gubernamental orientado al desarrollo de las "habilidades de pensamiento" de la infancia a la vejez, lo que estimuló el inicio de proyectos similares en muchos otros países (Collins y Mangieri, 1992; Greenwald, 1982, 1984; Herrnstein, Nickerson, Sánchez y Swets, 1986; Nickerson, 1988; Spitz, 1986; Sternberg, 1986).

En el capítulo 12 examinaremos la investigación realizada sobre los efectos que tienen en niños e infantes las intervenciones planeadas. Mientras tanto, también es importante advertir que cada vez se cuenta con más evidencias que demuestran que dichas intervenciones también son eficaces en etapas posteriores de la vida. Aunque a menor escala que los programas dirigidos a preescolares, los que se orientan a niños de edad escolar también han obtenido resultados alentadores (Bloom, 1976; Brown y Campione, 1986; Campione y Brown, 1987; Jacobs y Vandeventer, 1971; Olton y Crutchfield, 1969; Resnick y Glaser, 1976). Asimismo, los investigadores que trabajan con personas de más edad, como los estudiantes de universidades y escuelas profesionales, han traído progresos importantes a los instrumentos de aprovechamiento académico y de aptitud escolar (Bloom y Broder, 1950; Whimbey, 1975, 1977, 1980). La investigación con ancianos también ha producido evidencias de efectos de aprendizaje y transferencia después de las intervenciones de entrenamiento (Willis, Blieszner y Baltes, 1981); y otros investigadores que trabajan con niños y adolescentes con retardo mental han obtenido mejoras significativas (Babad y Budoff, 1974; Budoff y Corman, 1974; Feuerstein, 1980; Feuerstein *et al.*, 1987; Hamilton y Budoff, 1974; Rand, Tannenbaum y Feuerstein, 1979).[3]

Los programas citados brindan entrenamiento en habilidades cognoscitivas de gran aplicación, estrategias de resolución de problemas y hábitos de aprendizaje. De especial interés resulta el énfasis en la autosupervisión y la autocrítica como condiciones para un desempeño eficaz (Flavell, 1979; Owings, Petersen, Bransford, Morris y Stein, 1980; Whimbey, 1975). La evaluación que haga el individuo de su desempeño y el reconocimiento de lo que entiende o no son el primer paso hacia la mejora de su ejecución. Con demasiada frecuencia, el estudiante mediocre no logra diferenciar entre la verdadera comprensión y la comprensión inexacta o superficial. Se cuenta con evidencias de que los niños con problemas de aprendizaje son particularmente deficientes en la autocrítica y en la supervisión de su propia cognición (Kotsonis y Patterson, 1980).

[3] Para una evaluación crítica de la aproximación de Feuerstein, véase Anastasi (1980) y Blagg (1991).

En el capítulo 1 vimos algunos ejemplos de las habilidades cognoscitivas que se enseñan en los programas de desarrollo intelectual; ahí comparamos la capacitación en habilidades intelectuales de gran aplicación con la asesoría restringida a las preguntas de un examen. A ese respecto, vimos que una cuestión crucial por investigar en la evaluación de los programas de desarrollo intelectual es la magnitud y la duración de la transferencia o generalización de los resultados de la capacitación a nuevos contenidos y ambientes.

También debe considerarse el tiempo que requieren los niños mayores o los adultos para acumular el conocimiento que forma parte de la inteligencia y que contribuye a la disposición que muestra el sujeto para aprender material más avanzado. Hay evidencias de que, salvo en los niveles muy elementales, los conceptos y esquemas de solución de problemas se concretan a un dominio, es decir, se asocian con áreas temáticas particulares. Las habilidades en la resolución de problemas se relacionan estrechamente con el contenido organizado que el individuo ha acumulado en un campo específico de conocimiento (Bransford, Sherwood, Vye y Rieser, 1986; Brown y Campione, 1986; Glaser, 1984; Larkin, McDermott, Simon y Simon, 1980a; Neimark, 1987; Resnick y Neches, 1984; Richardson, Angle, Hasher, Logie y Stoltus, 1996). Aunque la persona mayor, que dispone de técnicas más eficaces de aprendizaje, puede almacenar el contenido con más rapidez de lo que podría haberlo hecho cuando niña, es poco realista esperar que así ocurra luego de un programa breve de entrenamiento, ya que entre mayor sea el individuo más grande será el vacío de conocimiento que debe ser llenado. No reconocer este hecho puede conducir a una decepción y a debilitar la confianza en la eficacia de los programas de capacitación.

MOTIVACIÓN E INTELIGENCIA

Aunque la clasificación de las pruebas en categorías resulta una costumbre conveniente, se debe tener claro que se trata de distinciones superficiales; por ejemplo, no se puede separar la personalidad de las aptitudes al interpretar las puntuaciones de las pruebas. El desempeño que muestra un individuo en un test de aptitud, en la escuela, en el trabajo o en cualquier otro contexto está influido por su motivación de logro, su persistencia, su sistema de valores, la existencia o ausencia de problemas emocionales discapacitantes y muchas otras características que tradicionalmente se clasifican bajo el rubro de "personalidad".

Cada vez se reconoce más la función que cumple la motivación del estudiante en el aprendizaje escolar (Bloom, 1976, capítulo 4; Budoff, 1987; Feuerstein *et al.*, 1987; J. G. Nichols, 1979; Renninger, Hidi y Krapp, 1992; R. E. Snow, 1989). Los intereses, las actitudes y el autoconcepto del individuo influyen en la disposición que muestra ante una tarea de aprendizaje, su deseo de aprenderla, la atención que preste al maestro y el tiempo que le dedica. También hay evidencias de que estas reacciones individuales se relacionan con el aprovechamiento académico (Baron, 1982; Dreger, 1968; J. McV. Hunt, 1981).

A un nivel más elemental, también es mayor el consenso respecto a que ya no es posible investigar las aptitudes independientemente de las variables afectivas (Anastasi, 1985b, 1994; Izard, Kagan y Zajonc, 1989; Kanfer, Ackerman y Cudeck, 1989;

parte IV; Moore e Isen, 1990; Saklofske y Zeidner, 1995; Salovey y Sluyter, 1997; R. E. Snow, 1992; Spaulding, 1994; Sternberg y Ruzgis, 1994).

Se conocen bien los efectos de los estados afectivos transitorios sobre el desempeño actual del individuo; pero también se sabe que es aún más importante el efecto acumulado de los rasgos de personalidad sobre la dirección y el alcance del desarrollo intelectual. Los datos correspondientes provienen de diferentes investigaciones, entre las que se incluyen algunos estudios longitudinales de largo plazo (Eichorn, Clausen, Hann, Honzik y Mussen, 1981) y, más recientemente, la aplicación del modelamiento de ecuaciones estructurales al análisis de las relaciones causales (Shavelson y Bolus, 1982). Estos estudios han arrojado evidencias que indican que la predicción del desarrollo intelectual de una persona puede mejorar considerablemente si se combina la información sobre su motivación y sus actitudes con las puntuaciones que haya obtenido en los tests de aptitud.

La motivación y otras variables afectivas contribuyen al desarrollo de las aptitudes al influir en el tiempo que el individuo dedica a una actividad en relación con el que dedica a otras ocupaciones que compiten por su atención. A partir de la investigación conducida durante 25 años sobre la motivación de logro, J. W. Atkinson y sus colaboradores (Atkinson, 1974; Atkinson, O'Malley y Lens, 1976) formularon un esquema que representa las relaciones entre habilidades, motivación y variables ambientales. Un concepto clave en este esquema es el tiempo dedicado a la tarea, es decir, el tiempo que el individuo destina a una actividad como el estudio o a realizar funciones relacionadas con el trabajo. La motivación influye tanto en la eficacia con la que se realiza una tarea como en el tiempo que le asigna en comparación con otras actividades. El nivel de desempeño depende tanto de las habilidades relevantes del individuo como de la competencia con la que las aplica a la tarea que enfrenta. El desempeño o producto final refleja la influencia combinada del nivel de aprovechamiento y el tiempo dedicado al trabajo.

Otro componente importante del esquema de Atkinson atañe al efecto acumulativo de la realización de la tarea sobre el propio desarrollo cognoscitivo y motivacional del individuo. Este elemento representa un circuito de retroalimentación de los rasgos de la persona, cuya influencia debe manifestarse en su desempeño futuro tanto en las pruebas como en las actividades de la vida cotidiana. El valor predictivo del esquema de Atkinson se demostró en simuladores computarizados y en el análisis empírico de datos longitudinales obtenidos con estudiantes de secundaria (Atkinson, 1974; Atkinson *et al.*, 1976; Lens, Atkinson y Yip, 1979).

El control de la atención resalta el efecto del tiempo dedicado a la tarea; es decir, lo que uno atiende, la profundidad y duración de la atención, contribuyen al propio crecimiento cognoscitivo. La selectividad de la atención conduce al aprendizaje selectivo, y habrá de diferir entre las personas expuestas a la misma situación. Más aún, este aprendizaje selectivo puede afectar el desarrollo relativo de diferentes aptitudes y, por consecuencia, contribuir a la formación de diversos patrones de rasgos (Anastasi, 1970, 1983a, 1986b). Básicamente, los diversos aspectos del control de la atención intensifican el efecto del tiempo que se dedica a las actividades importantes y por lo tanto aumentan su influencia sobre el desarrollo de aptitudes.

La relación entre personalidad e intelecto es recíproca: no sólo las características de la personalidad influyen en el desarrollo intelectual, sino que el nivel intelectual también puede repercutir en el desarrollo de la personalidad. Las evidencias de dicha relación provienen de los datos aportados por Plant y Minium (1967), estos investigadores seleccionaron al 25 por ciento superior e inferior de las muestras de cinco estudios longitudinales realizados con universitarios de acuerdo con las puntuaciones en tests de inteligencia. Luego, compararon los resultados de estos grupos contrastados en una serie de tests de personalidad aplicados a una o más de las muestras. Los instrumentos incluían mediciones de actitudes, valores, motivación y otros rasgos personales y no cognoscitivos. Los resultados del análisis revelaron una fuerte tendencia a que los grupos de mayor aptitud experimentaran más cambios sustanciales de personalidad "psicológicamente positivos" que los grupos de menor aptitud.

El éxito del individuo en el desarrollo y uso de sus aptitudes afecta su ajuste emocional, sus relaciones personales y su autoconcepto, punto éste en el que podemos ver más claramente la influencia recíproca de aptitudes y rasgos de personalidad. Existe una espiral continua en la que el desempeño del niño en la escuela, el patio de juegos y en otras situaciones contribuye a moldear su autoconcepto, que a su vez influye en su desempeño en cualquier etapa. A este respecto, el autoconcepto opera como una especie de profecía privada que se cumple.

En los años recientes se ha observado un aumento en el interés por el estudio de los factores afectivos en el desarrollo infantil. Diversas investigaciones han encontrado correlaciones importantes entre calificaciones de conducta infantil sobre variables de personalidad y el subsecuente desarrollo cognoscitivo evaluado por instrumentos como las escalas WISC-R y Stanford-Binet (Birns y Golden, 1972; R. B. McCall, 1976; Palisin, 1986; Yarrow y Pedersen, 1976). En general, y como resultado de sus experiencias tempranas, los infantes que en una situación de prueba muestran disposición a responder, sentimientos positivos y un interés activo tienen mayor probabilidad de aprender más, de avanzar más rápido en su desarrollo cognoscitivo y de responder más favorablemente a las actividades académicas que incluyen la interacción con adultos en tareas orientadas a las metas. Otra ventaja viene del efecto que la conducta del niño puede tener sobre la respuesta social de los adultos que se encargan de su cuidado, lo que a su vez aumentará las oportunidades de aprender del pequeño (Haviland, 1976; Wilson y Matheny, 1983).

Más específicamente, los estudios de la motivación que muestran los infantes para dominar su ambiente han revelado algunas relaciones prometedoras con la medición subsecuente de la competencia intelectual. En el comportamiento que muestra el niño al dominar su medio se incluyen la observación, la exploración y la manipulación de su ambiente. Por su propia naturaleza, este motivo debe contribuir en buena medida al desarrollo cognoscitivo, y de hecho cada vez son más las evidencias experimentales de que en efecto lo hace (Hrncir, Speller y West, 1985; White, 1978; Yarrow *et al.*, 1984; Yarrow *et al.*, 1983). De hecho, algunos de los datos obtenidos sugieren que los primeros indicadores de la motivación para dominar el ambiente pueden hacer una mejor predicción de la competencia intelectual posterior que las medidas tempranas. La investigación sobre la infancia conduce a un acercamiento entre el estudio del desarrollo afectivo y el cognoscitivo, y puede dar lugar a un uso integrado de ambos datos al interpretar los resultados de una prueba en cualquier nivel de edad.

ANÁLISIS FACTORIAL DE LA INTELIGENCIA

La investigación psicológica encaminada a la identificación de los rasgos mentales surgió del interés en la naturaleza y composición de la inteligencia humana.[4] Dicha investigación empieza al correlacionar las puntuaciones obtenidas por una muestra de personas en diversos tests de habilidad; posteriormente se somete la tabla de correlación a otros análisis matemáticos con el único propósito de identificar los factores o rasgos que los instrumentos tienen en común. Se conoce con el nombre genérico de *análisis factorial* a los distintos procedimientos de que actualmente se dispone para ese propósito.

La matriz factorial. El propósito principal del análisis factorial consiste en simplificar la descripción de los datos al reducir el número de variables o dimensiones necesarias. De modo que si encontramos que bastan cinco factores para dar cuenta de la varianza común de una batería de 20 pruebas, podemos sustituir las 20 puntuaciones originales con las cinco identificadas sin sacrificar información esencial. La práctica común consiste en conservar las puntuaciones originales que proporcionen la mejor medida de cada uno de los factores.

Todas las técnicas del análisis factorial empiezan con una tabla completa de las correlaciones encontradas en un conjunto de pruebas, a la que se conoce como matriz de correlación. El análisis termina con una matriz de factores, es decir, una tabla que muestra el peso o carga de cada factor en cada prueba. En la tabla 11.1 se muestra una matriz factorial hipotética que incluye dos factores, que se identifican en la parte superior, con sus pesos en cada una de las 10 pruebas en el renglón adecuado.

Se han derivado varios métodos diferentes para analizar un conjunto de variables en factores comunes. Ya en 1901 Pearson señaló el camino para realizar este análisis y

Tabla 11.1

Matriz factorial hipotética

Test	Factor I	Factor II
1. Vocabulario	.74	.54
2. Analogías	.64	.39
3. Completamiento de oraciones	.68	.43
4. Oraciones en desorden	.32	.23
5. Lectura de comprensión	.70	.50
6. Sumas	.22	−.51
7. Multiplicación	.40	−.50
8. Problemas aritméticos	.52	−.48
9. Relaciones de ecuaciones	.43	−.37
10. Completamiento de series numéricas	.32	−.25

[4] Para una revisión histórica, véase Anastasi (1984b).

Spearman (1904, 1927) elaboró un método precursor del análisis factorial moderno. Otras contribuciones al avance del método se deben a T. L. Kelley (1935) y Thurstone (1947b) en los Estados Unidos, y Burt (1941) en Inglaterra, aunque son muchos más quienes han desarrollado procedimientos alternativos, modificaciones y mejoras. El acceso a computadoras de alta velocidad ha permitido la adopción de técnicas más perfeccionadas y laboriosas. Aunque difieren en sus postulados iniciales, la mayor parte de estos métodos produce resultados similares. Para una introducción a los procedimientos específicos del análisis factorial, el lector puede consultar los textos de Comrey y Lee (1992) o Loehlin (1992), o bien puede encontrar un repaso general breve y sencillo de los conceptos y procedimientos básicos en Kim y Mueller (1978a, 1978b) y en P. Kline (1993).

Nuestro libro no pretende cubrir las bases matemáticas o los procedimientos de cálculo del análisis factorial; sin embargo, la comprensión de los resultados del análisis no tiene por qué limitarse a quienes dominan su metodología especializada, pues incluso sin saber cómo se calcularon las cargas factoriales es posible ver cómo se utiliza una matriz factorial para identificar e interpretar los factores. No obstante, la familiaridad con algunos otros términos y conceptos puede resultar de ayuda para una mejor lectura de los informes de la investigación sobre los análisis factoriales.

Los ejes de referencia. Se acostumbra representar los factores en forma geométrica, como ejes de referencia en términos de los cuales puede localizarse cada prueba, como

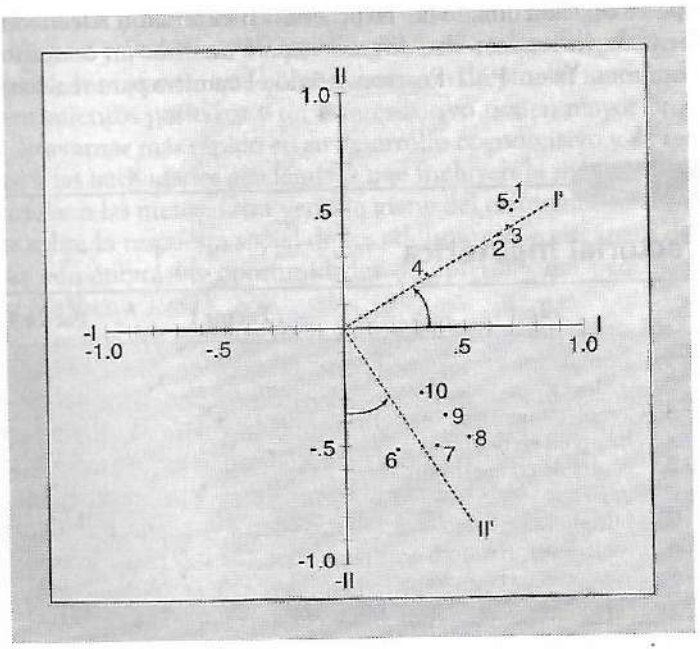


Figura 11.1. Modelo factorial hipotético que muestra los pesos de dos grupos de factores en 10 pruebas.

lo ilustra la figura 11.1. En la gráfica, cada una de las 10 pruebas de la tabla se sitúa según los dos factores que corresponden a los ejes I y II. Tenemos entonces que el punto que representa a la prueba uno se localiza al desplazarse una distancia de .74 a lo largo del eje I y una de .54 a lo largo del eje II. Los puntos que corresponden a las nueve pruebas restantes se localizan de la misma manera utilizando los pesos proporcionados en la tabla. Todos los pesos del factor I son positivos, pero observe que algunos de los pesos del factor II son negativos, lo que también se advierte en la figura 11.1, en la que las primeras cinco pruebas se agrupan en una parte de la gráfica y las cinco restantes en otra.

A este respecto, debe señalarse que los datos no determinan la posición de los ejes de referencia, y que la tabla original de correlación sólo determina la posición de las pruebas (es decir, los puntos en la figura 11.1) *en relación recíproca*. Si se cambian los ejes de referencia es posible localizar los mismos puntos, por lo que el analista factorial acostumbra rotar los ejes hasta obtener el patrón más satisfactorio y sencillo de interpretar. Se trata de un procedimiento legítimo que es similar a medir la longitud desde, digamos, Chicago, en lugar de hacerlo desde Greenwich.

En la figura 11.1 se rotaron los ejes de referencia a las posiciones I′ y II′ (representadas por las líneas punteadas);[5] para lograrlo, se siguieron los criterios de Thurstone de *múltiple positivo* y *estructura simple*. El primero requiere que los ejes giren hasta una posición que elimine todos los pesos negativos importantes, ya que la mayoría de los psicólogos considera que las cargas factoriales negativas son lógicamente inaplicables a los tests de aptitud, porque indican que entre mayor sea la calificación del individuo en el factor particular más pobre será su desempeño en la prueba. El criterio de estructura simple significa en esencia que cada instrumento tiene cargas en tan pocos factores como sea posible.[6] Ambos criterios permiten obtener factores cuya interpretación es sencilla y no ambigua. Si una prueba tiene una carga elevada en un factor y ninguna carga significativa en algún otro, podemos obtener información sobre la naturaleza del factor al examinar el contenido de la prueba; pero si ésta tiene cargas que van de moderadas a bajas en seis factores, es poco lo que puede decirnos sobre la naturaleza de cualquiera de los factores.

Puede verse que en los ejes rotados de la figura 11.1 todas las pruebas verbales (pruebas uno a cinco) caen en o cerca del eje I′ y que, de igual modo, las numéricas (pruebas de seis a diez) se agrupan alrededor del eje II′. En la tabla 11.2 se presentan las cargas de los nuevos factores, medidos a lo largo de los ejes rotados. El lector puede verificar fácilmente esas cargas factoriales preparando una "regla" en la que las unidades de la escala correspondan a la de la figura 11.1 que sirve para medir las distancias a lo largo de los ejes rotados. Las cargas factoriales de la tabla 11.2 no incluyen cifras negativas, salvo para cantidades insignificantes que pueden atribuirse a errores de muestreo. Todas las pruebas verbales tienen cargas elevadas en el factor I′ mientras que sus cargas en el factor II′ son prácticamente nulas. Por otro lado, las pruebas numéricas tienen cargas

[5] El lector puede pensar que el eje rotado II′ debería haberse denominado –II′ para corresponder al eje no rotado –II; sin embargo, la decisión de qué polo del eje se designa como positivo y cuál como negativo es una cuestión arbitraria. En el presente ejemplo, el eje rotado II′ ha sido "reflejado" para eliminar los pesos negativos.

[6] Este criterio requiere que las pruebas tengan cargas en algunos factores que no difieran significativamente de cero. Ese requisito puede ahora probarse en forma empírica por medio de procedimientos estadísticos para encontrar el error estándar de las cargas factoriales (Cudeck y O'Dell, 1994).

Tabla 11.2

Matriz factorial rotada

Prueba	Factor I′	Factor II′
1. Vocabulario	.91	−.06
2. Analogías	.75	.02
3. Completamiento de oraciones	.80	.00
4. Oraciones en desorden	.39	−.02
5. Lectura de comprensión	.86	−.04
6. Sumas	−.09	.55
7. Multiplicación	.07	.64
8. Problemas aritméticos	.18	.68
9. Relaciones de ecuaciones	.16	.54
10. Completamiento de series numéricas	.13	.38

(Datos de la figura 11.1)

altas en el factor II′ y cargas insignificantes en el factor I′. La rotación de los ejes de referencia facilita de esta manera la identificación y la denominación de ambos factores así como la descripción de la composición factorial. En la práctica, el número de factores suele ser mayor de dos, una condición que si bien complica la representación geométrica y el análisis estadístico, no altera el procedimiento básico.

Algunos analistas factoriales utilizan un modelo teórico para dirigir la rotación de los ejes y también toman en consideración la invarianza o la corroboración de los mismos factores en investigaciones independientes pero comparables. En la actualidad, el uso del análisis factorial es mayor en una función confirmatoria que en una exploratoria, y a menudo se combina con el modelamiento de ecuaciones estructurales (véase el capítulo 5) para evaluar un modelo formulado teóricamente de la operación de distintas variables en la ejecución de una tarea (véase, por ejemplo, Loehlin, 1992).

Interpretación de los factores. Una vez que se ha calculado la matriz de factores rotados, podemos proceder a interpretar y dar nombre a los factores. Este paso requiere más de discernimiento psicológico que de entrenamiento estadístico. Para entender la naturaleza de un factor en particular, podemos examinar sencillamente las pruebas con mayores cargas en el factor y tratar de descubrir qué procesos psicológicos tienen en común. Entre más sean los instrumentos con cargas elevadas en determinado factor, mejor podemos definir su naturaleza; por ejemplo, en la tabla 11.2 resulta evidente que el factor I′ es verbal y que el factor II′ es numérico.

Las cargas factoriales presentadas en la tabla 11.2 también representan la correlación de cada prueba con el factor.[7] Al respecto debe recordarse que esta correlación representa la validez factorial de la prueba (capítulo 5). Así, a partir de la tabla 11.2 podemos

[7] Esto es así únicamente cuando se realiza una rotación ortogonal de los ejes. Con la rotación oblicua, que definimos en otra sección del capítulo, las cargas y las correlaciones factoriales mantienen una relación simple y cada una puede encontrarse a partir de la otra por medio de los cálculos apropiados.

decir que la validez factorial de la prueba de vocabulario como medida del factor verbal es .91 y que la validez factorial de la prueba de sumas, en términos del factor numérico, es de .55. Desde luego, la validez de los primeros cinco instrumentos como medidas del factor numérico es insignificante, mientras que las últimas cinco prácticamente no tienen validez como medidas del factor verbal.

Composición factorial de una prueba. Uno de los teoremas básicos del análisis factorial afirma que la varianza total de la prueba es la suma de las varianzas aportadas por los factores comunes (compartidos con otros instrumentos), la varianza correspondiente a los factores específicos (que sólo existen en la prueba) y la varianza de error. En el capítulo 4 aprendimos a encontrar la varianza de error en el análisis de las puntuaciones de las pruebas; por ejemplo, si el coeficiente de confiabilidad de un instrumento es .83, podemos concluir que 17 por ciento de la varianza de las puntuaciones de la prueba corresponde a la varianza de error (1.00 – .83 = .17). Además, mediante el análisis factorial podemos subdividir las fuentes de varianza que contribuyen al desempeño en cualquier prueba.

Consideremos las dos pruebas hipotéticas presentadas en la tabla 11.3. Para cada una, la tabla proporciona su carga factorial en los factores verbal (V), numérico (N) y de razonamiento (R), así como su coeficiente de confiabilidad. Como la carga factorial también representa la correlación entre la prueba y el factor, el cuadrado de la carga factorial nos proporciona la varianza común entre el instrumento y el factor. En la última parte de la tabla, se ha elevado al cuadrado cada carga factorial para mostrar la contribución proporcional del factor a la varianza total de las puntuaciones obtenidas en la prueba. Así, encontramos que en la prueba de razonamiento aritmético, 16 por ciento de la varianza puede atribuirse al factor verbal, 30 por ciento al factor numérico y 36 por ciento al factor de razonamiento. La varianza de error en la última columna se encuentra al restar el coeficiente de confiabilidad de la varianza total (1.00 – 90 = .10). Lo que quede representa la especificidad de esta prueba, es decir, la proporción "verdadera" de la varianza que no comparte con ninguna otra prueba en la que se analizó el factor. Para la prueba de razonamiento aritmético tenemos:

$$.16 + .30 + .36 + .10 = .92$$
$$1.00 - .92 = .08$$

La figura 11.2 presenta una descomposición gráfica de las fuentes de varianza para las dos pruebas de la tabla 11.3.

La ejecución de cualquier individuo en esos dos instrumentos depende de la cantidad que posea en cada una de las habilidades o factores relevantes, así como de los pesos relativos que éstos tengan en el instrumento utilizado. Por ende, si conocemos la puntuación del individuo en los factores V, N y R, expresados en las mismas unidades, podemos pesar cada puntuación multiplicándola por la correspondiente carga factorial. Al sumar esos productos obtendremos una estimación de la puntuación del individuo en la prueba, cuya exactitud aumentará conforme vaya disminuyendo la contribución de factores específicos y de factores de error al instrumento.

Siguiendo el ejemplo de la tabla 11.3, si un sujeto obtiene una calificación muy alta en V, puede ayudarle más en la prueba de razonamiento aritmético que en la de

Tabla 11.3

Fuente de varianza de las puntuaciones de la prueba

Prueba	Cargas factoriales comunes			Coeficiente de confiabilidad	Contribución proporcional				
	V	N	R		V	N	R	*Específica*	*Error*
1. Razonamiento aritmético	.40	.55	.60	.90	.16	.30	.36	.08	.10
2. Multiplicación	.10	.70	.30	.85	.01	.49	.09	.26	.15

multiplicación; de hecho, le ayudaría cuatro veces más, ya que el peso de V es cuatro veces mayor en el razonamiento aritmético que en la multiplicación (.40 contra .10). De los tres factores comunes, N tiene el mayor efecto sobre la multiplicación (carga = .70) y R en razonamiento aritmético (carga = .60).

Cargas factoriales y correlación. El segundo teorema del análisis factorial tiene que ver con la relación entre las cargas factoriales y las correlaciones entre variables. La correlación entre dos variables cualesquiera es igual a la suma de los productos cruzados de las cargas de sus factores comunes. Los factores específicos y de error no pueden contribuir a la correlación entre variables porque son particulares de cada variable, de ahí que la correlación entre dos variables dependa únicamente de los factores que

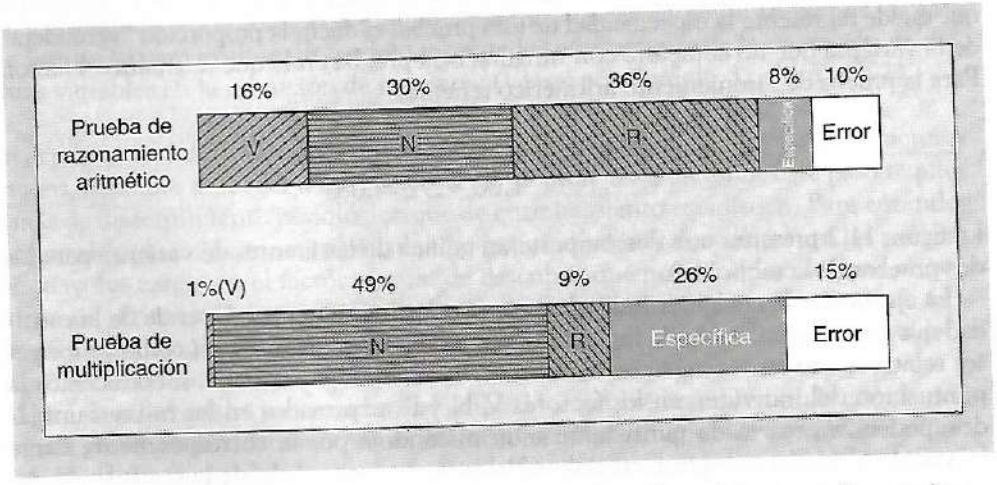


Figura 11.2. Porcentaje de varianza común, específica y de error en dos pruebas hipotéticas.

(Datos de la tabla 11.3)

comparten. Entre mayor sea el peso de los factores comunes, mayor será la correlación entre ellas. Como se muestra a continuación, para encontrar la correlación entre los dos instrumentos de la tabla 11.3 se multiplican las cargas de los tres factores comunes en las dos pruebas y se suman los productos:

$$r_{12} = (.40)(.10) + (.55)(.70) + (.60)(.30) = .60$$

Ejes oblicuos y factores de segundo orden. Los ejes de la figura 11.1 se conocen como *ejes ortogonales* porque forman ángulos rectos. Ocasionalmente los grupos de pruebas están tan localizados que es posible obtener un mejor ajuste con *ejes oblicuos*, en cuyo caso los factores estarían correlacionados. Algunos investigadores sostienen que siempre se debería utilizar factores ortogonales, o no correlacionados, porque proporcionan una imagen más clara y sencilla de las relaciones entre rasgos; otros insisten en usar los ejes oblicuos cuando se ajustan mejor a los datos porque afirman que la mayor parte de las categorías significativas no tiene por qué no estar correlacionada. Un ejemplo evidente es el del peso y la estatura que, a pesar de su elevada correlación, han demostrado ser categorías útiles en la medición del físico.

Cuando los factores están correlacionados es posible someter sus correlaciones al mismo análisis estadístico que empleamos con las correlaciones entre pruebas. En otras palabras, podemos "factorizar los factores" y derivar *factores de segundo orden*. Este proceso se ha seguido en varios estudios con variables de aptitud y de personalidad. Algunas investigaciones con tests de aptitud han dado lugar a un solo factor general de segundo orden. Como regla, los investigadores estadounidenses intentan explicar toda varianza común que sea posible mediante grupos de factores, y, si los datos lo justifican, luego tratan de identificar un factor general como factor de segundo orden. Por su parte, los psicólogos ingleses suelen comenzar con un factor general al que atribuyen la mayor parte de la varianza común y luego pasan a los factores de grupo para explicar cualquier correlación residual. Esas diferencias de procedimiento reflejan distinciones en las orientaciones teóricas, las que veremos en la siguiente sección.

TEORÍAS DE LA ORGANIZACIÓN DE LOS RASGOS

Durante más de medio siglo se han aplicado las técnicas estadísticas del análisis factorial para tratar de comprender la naturaleza y organización de las habilidades implicadas en una gran diversidad de actividades humanas. No obstante, esas técnicas aún se asocian principalmente con la investigación sobre habilidades cognoscitivas o "inteligencia", en las que tuvo su origen el análisis factorial. Recientemente concluyó un análisis de los estudios disponibles del análisis factorial de las habilidades cognoscitivas, que ha dado por resultado un resumen impresionante que abarca más de 70 años de investigación (Carroll, 1993). La empresa representa mucho más que una simple revisión de la bibliografía, ya que incluye el nuevo análisis de más de 450 conjuntos de datos de los estudios originales. En la misma fuente se describen y evalúan en su marco histórico diferentes modelos teóricos de la inteligencia. En esta sección, sólo consi-

deraremos algunas de las teorías más conocidas, seleccionadas por su impacto sobre la elaboración y uso de las pruebas.

La teoría bifactorial. La primera teoría sobre la organización de los rasgos basada en un análisis estadístico de las puntuaciones de una prueba fue la teoría bifactorial del psicólogo inglés Charles Spearman (1904, 1927). En su planteamiento original, la teoría sostenía que todas las actividades intelectuales comparten un único factor común, llamado *factor general* o *g*. Postulaba además la existencia de numerosos factores *específicos* o *s*, cada uno de los cuales se dirigía concretamente a una actividad; en consecuencia, la correlación positiva entre dos funciones cualesquiera se atribuía al factor *g*, es decir, entre más "saturadas" estuvieran dos funciones con *g*, mayor sería la correlación entre ellas; por otro lado, la presencia de condiciones específicas tendería a disminuir la correlación entre las funciones.

Aunque la teoría propone dos tipos de factores (general y específicos), sólo *g* da cuenta de la correlación, por lo que, a diferencia de otras teorías sobre las relaciones entre rasgos, y aunque haya persistido el nombre original, podría caracterizarse de manera más precisa como una teoría de factor único. La figura 11.3 ilustra, de acuerdo con la teoría, la base de la correlación entre instrumentos. En este ejemplo y según se aprecia en las áreas oscuras, las dos primeras pruebas tienen una correlación alta ya que muestran una elevada saturación con *g*. Por su parte, la tercera prueba tendría una correlación baja con las otras dos porque contiene muy poca *g*. En cada instrumento, el área blanca representa la varianza específica y la varianza de error.

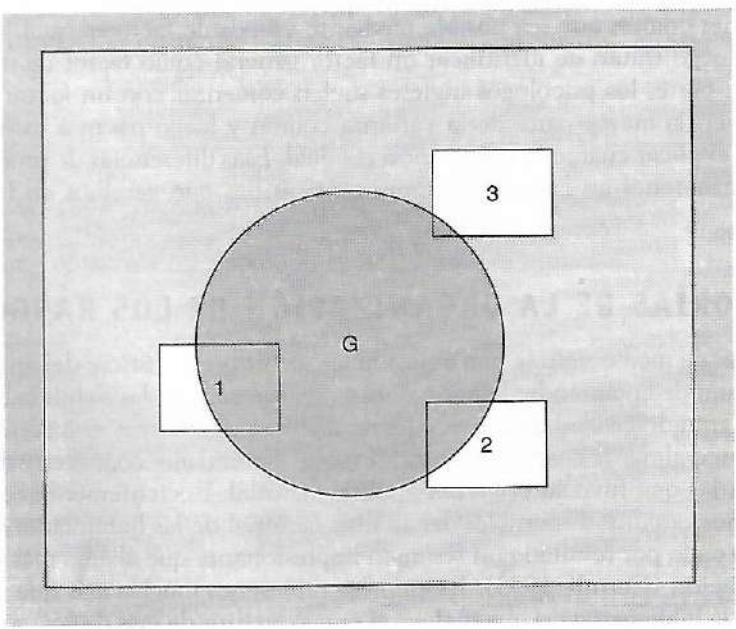


Figura 11.3. Modelo de correlación que fundamenta la teoría de los dos factores.

De acuerdo con la teoría bifactorial, el propósito de la evaluación psicológica debería ser la medición de la cantidad de g que posee cada individuo, ya que al encontrarse en todas las habilidades, es la única base para predecir la ejecución del individuo en distintas situaciones. Sería inútil medir los factores específicos, ya que, por definición, cada uno opera en una única actividad. En consecuencia, Spearman propuso sustituir la heterogénea colección de reactivos que se encuentran en los tests de inteligencia con una sola prueba que tuviera una elevada saturación de g. Además, sugería utilizar los instrumentos que evalúan las relaciones abstractas por considerar que son los que pueden hacer una mejor medición de g. Entre los instrumentos construidos como medidas de g se encuentran las Matrices Progresivas de Raven y el Test de Inteligencia Culturalmente justo de Cattell.

Desde el inicio, Spearman se percató de que la teoría bifactorial debería ser calificada. Cuando se comparan actividades similares, parte de la correlación entre ellas puede estar por encima de lo atribuible al factor g, lo que implica que, además de los factores general y específicos, debe existir otra clase intermedia que no es tan universal como g ni tan específica como s. Dicho factor, común a un grupo de actividades pero no a todas, se designó como *factor de grupo*. En la primera formulación de su teoría, Spearman admitía la posibilidad de que existieran factores de grupo muy pequeños e insignificantes; posteriormente, y de acuerdo con las investigaciones de varios de sus estudiantes, incluyó factores de grupo mucho más amplios, como las habilidades aritmética, mecánica y lingüística.

Teorías de factores múltiples. La postura estadounidense ante la organización de los rasgos, basada en la investigación previa del análisis factorial, reconocía la existencia de factores de grupo moderadamente amplios, cada uno de los cuales podía contribuir con diferentes pesos a distintas pruebas; por ejemplo, un factor verbal tendría un peso elevado en una prueba de vocabulario, menor en una de analogías verbales y más pequeño aún en una de razonamiento aritmético. La figura 11.4 ilustra las correlaciones entre cinco pruebas en términos de un modelo de factores múltiples. Las que tienen las pruebas uno, dos y tres son resultado de sus cargas comunes con el factor verbal (V); de igual modo, la correlación entre las pruebas tres y cinco es resultado del factor espacial (E), mientras que el factor numérico (N) es responsable de la correlación entre las pruebas cuatro y cinco. La composición factorial de las pruebas tres y cinco es compleja porque las dos tienen cargas apreciables en más de un factor: V y E en la prueba tres y N y E en la prueba cinco. El segundo teorema básico del análisis factorial, que estudiamos en la sección anterior, nos permite decir algo acerca de la magnitud relativa de las correlaciones; por ejemplo, la prueba tres tendrá una correlación más alta con la prueba cinco que con la dos porque el peso del factor E en las pruebas tres y cinco (las áreas con líneas diagonales) es mayor que el peso de V en las pruebas dos y tres (las áreas con líneas horizontales).

La publicación de la obra de T. L. Kelley *Crossroads in the Mind of Man* (1928) preparó el camino para la realización de un gran número de estudios que buscaban factores de grupo particulares. Entre los principales factores que propuso se encuentran la manipulación de las relaciones espaciales, la facilidad con los números y con el material verbal, la memoria y la velocidad. Más adelante, otros investigadores modificaron y extendieron la lista con la ayuda de los modernos métodos de análisis factorial.

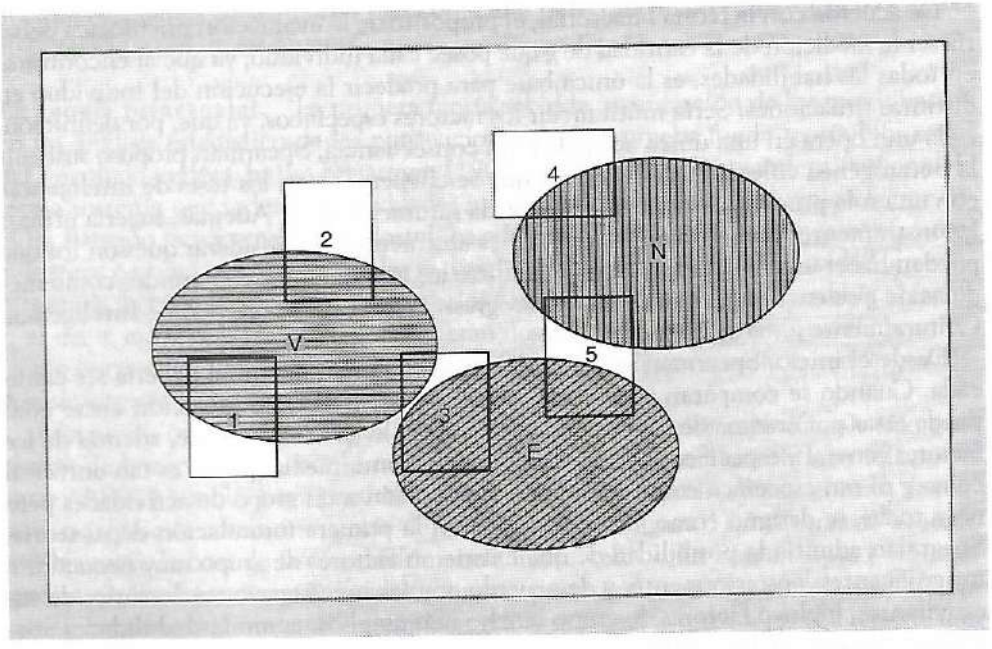


Figura 11.4. Modelo de correlación que fundamenta las teorías de factores múltiples.

Thurstone fue uno de los principales exponentes de la teoría de factores múltiples. Sobre la base de la investigación realizada por él y por sus estudiantes postuló alrededor de una docena de factores de grupo a los que denominó como "habilidades mentales primarias"; de éstas, entre las corroboradas con más frecuencia por el trabajo del propio Thurstone y de otros investigadores independientes (French, 1951; Harman, 1975; Thurstone, 1938; Thurstone y Thurstone, 1941) se encuentran las siguientes:

V. *Comprensión verbal*. El factor principal en pruebas como las de comprensión de lectura, analogías verbales, oraciones en desorden, razonamiento verbal e igualación de proverbios. Las pruebas de vocabulario lo miden más adecuadamente.

F. *Fluidez verbal*. Se encuentra en pruebas como los anagramas, rimas o nombrar palabras de una determinada categoría (por ejemplo, nombres de niños o palabras que empiecen con la letra T).

N. *Números*. Se identifica más con la rapidez y la precisión de los cálculos aritméticos simples.

S. *Espacio*. Puede representar dos factores distintos: uno cubre la percepción de relaciones espaciales o geométricas fijas; el otro, visualizaciones manipulatorias de transformaciones o cambio de posiciones (McGee, 1979; Portegal, 1982).

M. *Memoria asociativa*. Se encuentra principalmente en las pruebas que exigen la memorización de pares asociados. Hay evidencias que indican que este factor puede reflejar el grado en el que se utilizan soportes de memoria (Christal, 1958). Las evidencias desmienten la presencia de un factor más general en todas

las pruebas de memoria. Algunas investigaciones indican la existencia de otros factores restringidos de memoria, como la memoria para las secuencias temporales y para la posición espacial.

P. *Velocidad perceptual*. Captar de manera rápida y precisa las semejanzas, diferencias y los detalles visuales. Éste es uno de varios factores identificados en las tareas perceptuales (Thurstone, 1944) y puede tratarse del mismo factor de velocidad señalado por Kelley y otros investigadores.

I. (*o R*). *Inducción* (o *razonamiento general*). La identificación de este factor fue menos clara. Originalmente, Thurstone propuso un factor inductivo y uno deductivo. El último se medía mejor con pruebas de razonamiento silogístico y el primero con instrumentos que requerían que el examinado encontrara una regla, como en un tests de completamiento de series de números; sin embargo, las evidencias a favor del factor deductivo resultaron mucho más débiles que para el inductivo. Además, otros investigadores postularon la existencia de un factor de razonamiento general, que se mide en forma más adecuada con pruebas de razonamiento aritmético.

Debe advertirse que la distinción entre factores generales, de grupo y específicos no es tan sencilla como parece. Si el número o la variedad de pruebas que incluye una batería es pequeño, un solo factor general puede dar cuenta de todas las correlaciones. Pero cuando los mismos instrumentos se incluyen en una batería más grande que contiene un conjunto más heterogéneo de pruebas, el factor general original puede aparecer como factor de grupo, común a algunas de las pruebas pero no a todas. De modo similar, puede ocurrir que una única prueba de la batería original represente cierto factor, pero que varios instrumentos de la batería más grande lo compartan. Tal factor, que en la batería original puede haberse identificado como específico, podría resultar de grupo en la batería más amplia. A la luz de esas consideraciones, no resulta sorprendente que las investigaciones factoriales realizadas en áreas especiales hayan dado lugar a muchos factores en lugar de las pocas capacidades mentales primarias que se habían señalado en cada área. Así ha sucedido en los estudios de las pruebas verbales, perceptuales, de memoria y de razonamiento.

La investigación factorial parece haber producido una sorprendente multiplicación de factores. A la fecha, el número de factores cognoscitivos postulados por distintos investigadores supera los 100. Se alcanzó cierto orden con la identificación cruzada de factores informados por diferentes investigadores que a menudo empleaban nombres distintos (Ekstrom, French y Harman, 1979; French, 1951; Harman, 1975). La identificación cruzada se logra cuando las investigaciones comparadas utilizan varias pruebas en común. Para facilitar este proceso, un grupo de analistas preparó un juego de "tests de referencia" para medir los principales factores de aptitud identificados en esa época. Este juego, distribuido por el Servicio de Pruebas Educativas (*Educational Testing Service*, Ekstrom, French, Harman y Dermen, 1976; ETS *kit*, 1976), permite a los investigadores planear la investigación factorial de modo que puedan incluir en sus baterías algunas pruebas comunes.

Es evidente que el número de factores sigue siendo grande a pesar de esos esfuerzos de simplificación y coordinación. La conducta humana es variada y compleja, por lo que no es realista esperar que alrededor de una docena de factores la describa en forma adecuada; sin embargo, para ciertos propósitos podemos elegir los factores que sean apropia-

dos a su naturaleza y amplitud. Así, si queremos elegir entre los solicitantes para un trabajo mecánico y muy especializado, querremos medir los factores espaciales y perceptuales que corresponden a los requisitos del trabajo. Por otro lado, si lo que queremos es seleccionar a los estudiantes que ingresan a la universidad, pueden ser factores más relevantes como la comprensión verbal, la facilidad numérica y el razonamiento general.

Modelo de la estructura del intelecto. Algunos investigadores han intentado simplificar la imagen de las relaciones entre los rasgos, organizándolos en un esquema sistemático. Guilford (1967, 1988; Guilford y Hoepfner, 1971) se apoyó en un programa de investigación conducido durante más de dos décadas para proponer un esquema similar a una caja, al que denominó modelo de la estructura del intelecto (EI). En la figura 11.5 se ilustra este modelo[8] que clasifica los rasgos intelectuales en tres dimensiones:

Operaciones. Corresponde a lo que hace el individuo. Incluye la cognición, el registro y la retención en memoria, la producción divergente (que es importante en la actividad creativa), la producción convergente y la evaluación.

Contenidos. Atañe a la naturaleza de los materiales o información sobre la que se realizan las operaciones. Incluye contenidos visual, auditivo, simbólico (letras, números), semántico (palabras) y conductual (información acerca de la conducta, actitudes y necesidades de otra persona).

Productos. Corresponde a la forma en que el individuo procesa la información. Los productos se clasifican en unidades, clases, relaciones, sistemas, transformaciones e implicaciones.

El modelo contiene 180 celdillas porque la clasificación incluye $6 \times 5 \times 6$ categorías. En cada celdilla se espera al menos un factor o habilidad, aunque algunas pueden contener más de uno. Cada factor se describe en términos de las tres dimensiones. Al concluir el Proyecto de Investigación de Aptitudes, un programa de investigación sobre el modelo de EI conducido durante 20 años, Guilford y sus colaboradores pudieron identificar 98 de los factores anticipados (Guilford y Hoepfner, 1971). En el libro de Guilford y Hoepfner (1971, apéndice B) se encuentra una lista de las pruebas elaboradas durante el proyecto, presentadas por orden alfabético y acompañadas por una breve descripción.

Aunque al menos una prueba se diseñó explícitamente de acuerdo con el modelo EI, el Test de las Habilidades de Aprendizaje de Estructura del Intelecto (*Structure of Intellect Learning Abilities Test*, Meeker, Meeker y Roid, 1985), ha sido poco el impacto del modelo en el desarrollo y la aplicación de pruebas para uso general.[9] Recordemos que el modelo EI, como todos los modelos sobre la organización de los rasgos, proporciona un esquema para representar las correlaciones obtenidas entre variables. Debido al método de rotación de ejes, la corroboración empírica del modelo EI no excluye otros modelos (Carroll, 1972; Horn y Knapp, 1973), por lo que es posible encontrar que diferentes rotaciones de los mismos datos muestren un ajuste igualmente bueno con otros modelos. De hecho, al volver a analizar los datos originales de Guilford se

[8] Como resultado de las investigaciones posteriores de Guilford (Guilford, 1967), este modelo difiere del anterior en el hecho de que divide la faceta de contenido figurativo en componentes visual y auditivo, y la faceta de memoria de operación en los componentes de registro y retención (Guilford, 1988).

[9] Para referencias relevantes, véase Carroll (1993), especialmente pp. 57-60.

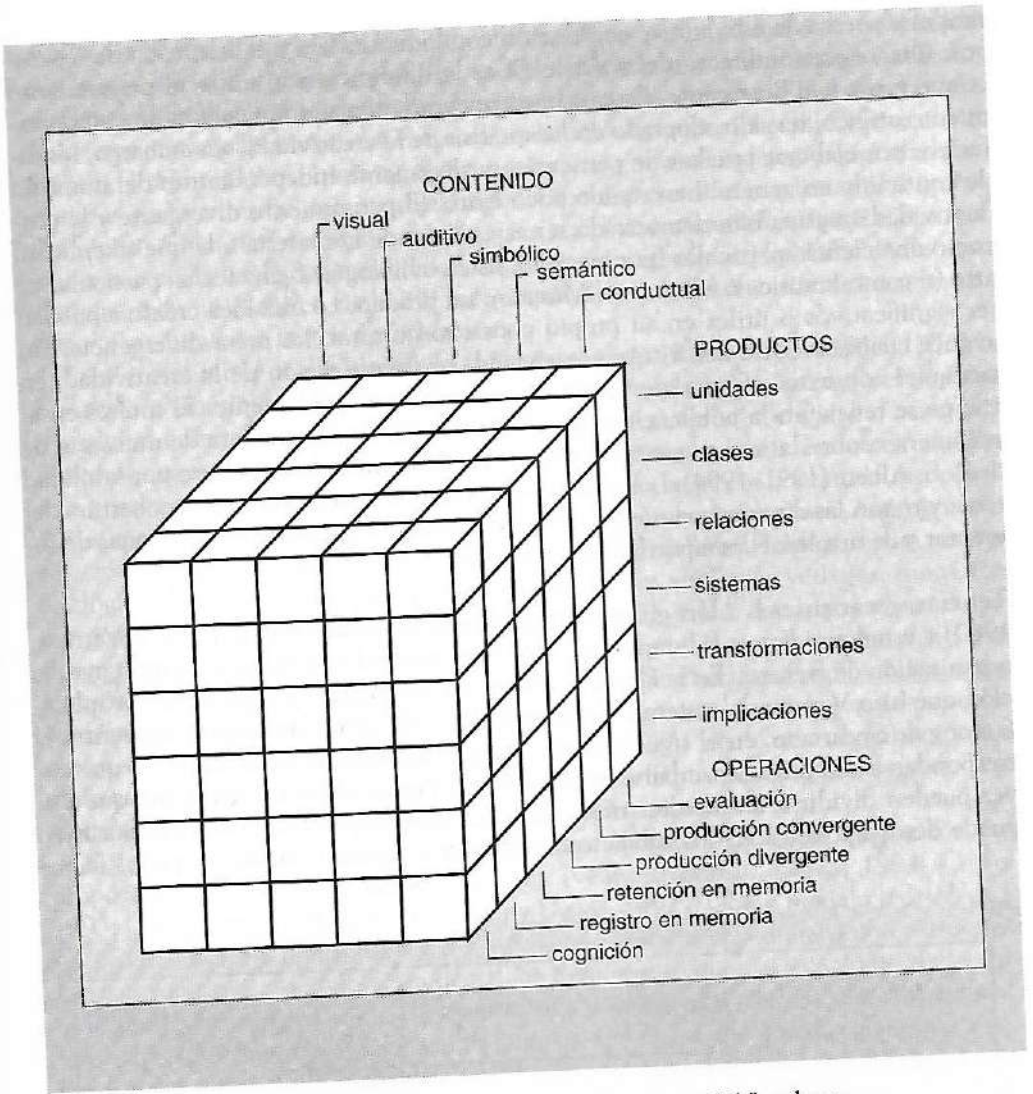


Figura 11.5. Modelo tridimensional de la Estructura del Intelecto.

(Revisión de Guilford, 1988, p. 3. Copyright © 1988 por *Educational and Psychological Measurement*. Reproducido con autorización).

ha descubierto que otros modelos tienen un mejor ajuste a los datos que el modelo EI y que son, además, más congruentes con las interpretaciones teóricas y prácticas (Bachelor, 1989; Carroll, 1993).

Por otro lado, un efecto indirecto del proyecto EI fue que, al identificar los factores, dirigió la atención a la distinción entre operación y contenido, lo que ayudó a aclarar los factores identificados mediante el análisis factorial, los procesos investigados por la psicología cognoscitiva y la relación entre ambos. En la última parte del capítulo vere-

mos más acerca de este punto, en relación con la naturaleza y el desarrollo de los rasgos. Otro efecto indirecto del modelo EI es la diferenciación entre el pensamiento convergente y el divergente. Este último concepto, con su implicación de comportamiento atípico, ha sido adoptado en los análisis de la creatividad; sin embargo, los esfuerzos por elaborar pruebas de pensamiento divergente independientes del dominio de contenido en general han tenido poco éxito. El pensamiento divergente y la productividad creativa han demostrado ser específicos de la materia a la que se aplican, como una ciencia particular (por ejemplo, física o biología) o una forma particular de arte (digamos, música o escultura). Más aún, un producto o una idea creativa pueden ser significativos o útiles en su propio contexto cultural. La mera divergencia, sin avance cualitativo, no constituye creatividad. La investigación de la creatividad, en cualquier contexto o forma que se manifieste, progresa rápidamente. El interés en el campo se refleja en la publicación durante la década de los noventa de una serie de volúmenes sobre la investigación de la creatividad realizada bajo la responsabilidad de R. S. Albert (1991–1994). Los volúmenes de la serie hacen una amplia cobertura del tema y tratan las diversas manifestaciones de la creatividad así como las formas de fomentar y desarrollar el comportamiento creativo.

Teorías jerárquicas. Un grupo de psicólogos británicos (Burt, 1949; y Vernon, 1960) y estadounidenses (Humphreys, 1962) propuso un esquema alternativo para la organización de factores. En la figura 11.6 se presenta el diagrama que ilustra la aplicación que hizo Vernon del sistema. En la parte superior de la jerarquía se encuentra el factor *g* de Spearman, en el siguiente nivel se localizan dos factores de grupo que corresponden a las aptitudes verbal-educativa (*v:ed*) y práctica-mecánica (*p:m*), que a su vez pueden dividirse en factores más finos; por ejemplo, el factor verbal-educativo puede descomponerse en los subfactores verbal y numérico, mientras que el factor

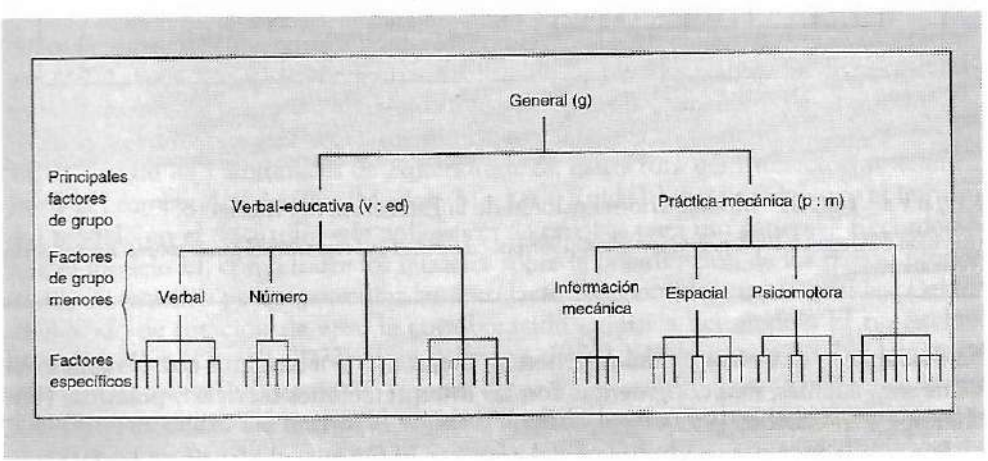


Figura 11.6. Modelo de una organización jerárquica de habilidades.

(Adaptado de P. E. Vernon, 1960, p. 22. Copyright © 1960, Methuen and Co., Ltd. Reproducido con autorización.)

práctico-mecánico puede dividirse en subfactores como habilidades de información mecánica, espacial y psicomotora. Es posible identificar subfactores aún más finos mediante el análisis subsecuente de, digamos, las tareas verbales. En el nivel inferior de la jerarquía se encuentran los factores específicos. En una elaboración posterior del modelo, Vernon (1969) incluyó algunas relaciones más complejas y contribuciones cruzadas de factores del tercer nivel, especialmente los relacionados con el aprovechamiento educativo y vocacional. Así, las habilidades científicas y técnicas están asociadas con la habilidad espacial y la información mecánica; las habilidades matemáticas están asociadas a las habilidades espacial y numérica y más directamente al factor g mediante un factor de inducción.

Para enfrentar la proliferación de factores, también Humphreys (1962, 1970) recomendó un modelo jerárquico, pero en lugar de considerar un único nivel de factores, sugirió que cada individuo que elaborara o usara una prueba eligiera el nivel de la jerarquía que resultara más apropiado para sus propósitos. Además, reconoció que una sola prueba puede clasificarse en más de una jerarquía, según su contenido, procesos y otras facetas. Para medir cualquier faceta, propuso que la prueba fuera *heterogénea* en relación con las otras; por ejemplo, si nos interesa la habilidad del sujeto para resolver analogías, deberíamos usar una prueba que incluyera analogías verbales, numéricas, pictóricas y espaciales; si deseamos medir su habilidad verbal, deberíamos hacerlo con diferentes reactivos, como los de vocabulario, analogías y completamiento de series. Este procedimiento contrasta con el empleado por Guilford, quien pretendía separar los factores (y pruebas) para cada celdilla homogénea en su clasificación de tres direcciones. Con todo, en su último trabajo Guilford (1981) siguió una organización parcialmente jerárquica al identificar los factores de orden superior entre algunos de los factores de su modelo original de la estructura del intelecto.

El modelo jerárquico de la inteligencia recibe cada vez mayor aceptación para propósitos tanto teóricos como prácticos (véase Anastasi, 1992a, 1994; Carroll, 1993; Gustafsson, 1984; 1989; Lubinski y Dawis, 1992). Como modelo teórico de las relaciones entre los rasgos, ajusta el factor general (la g de Spearman) con patrones de factores múltiples. Metodológicamente, ha demostrado que las soluciones de factores múltiples y la jerárquica son de manera matemática equivalentes e intercambiables (Harman, 1976, capítulo 15; Schmid y Leiman, 1957). La solución de ejes oblicuos (con factores correlacionados) que da lugar al modelo jerárquico puede convertirse en una solución de ejes ortogonales (con factores no correlacionados). En la última solución, aparecen factores de segundo orden como de primer orden de diferente amplitud. Los factores más amplios tienen cargas en más variables que los factores más restringidos.

Desde un punto de vista práctico, una ventaja de las pruebas desarrolladas a partir de un patrón jerárquico es que combinan una cobertura amplia con la flexibilidad del uso. Para adecuarse a diferentes objetivos, el usuario puede elegir una calificación global de la batería, o una o más calificaciones de grupos de pruebas que miden factores más definidos. En ciertas circunstancias pueden utilizarse las calificaciones de subtests individuales para identificar los puntos débiles o fuertes en ciertas habilidades especializadas.

En el capítulo 2 hablamos del creciente impacto del modelo jerárquico en la elaboración y el uso de pruebas de habilidad como un desarrollo histórico que permitió unificar los tests tradicionales de inteligencia con las baterías de aptitudes múltiples. Asimismo,

en el capítulo 8 vimos algunos ejemplos de la aproximación jerárquica a las pruebas individuales y en el capítulo 10 a las pruebas colectivas. Entre los ejemplos más claros de la combinación de una cobertura amplia y un uso flexible se encuentran las Escalas de Habilidad Diferencial (Elliott, 1990b) para la evaluación individual y la Batería de Aptitudes Multidimensionales (Jackson, 1994b) para la evaluación colectiva. Respecto a la flexibilidad de los niveles de calificación de esas pruebas, una característica notable es el hecho de que proporcionan normas para interpretar las calificaciones en el nivel elegido. Sin las normas de niveles separados sería imposible obtener una evaluación apropiada del desempeño del examinado en diferentes niveles de la jerarquía.

NATURALEZA Y DESARROLLO DE LOS RASGOS

El hecho de que diferentes investigadores puedan obtener modelos tan disímiles de la organización de los rasgos resulta menos sorprendente si comprendemos que los rasgos identificados mediante el análisis factorial son sencillamente la expresión de la correlación entre mediciones de conducta, es decir, que no se trata de entidades subyacentes o factores causales, sino de categorías descriptivas, de ahí que sea posible aplicar diferentes principios de clasificación a los mismos datos. El concepto de factores como categorías descriptivas se hizo explícito en los primeros escritos de Thomson (1948), Burt (1941, 1944) y Vernon (1960) en Inglaterra, y en los de R. C. Tryon (1935) en los Estados Unidos, quienes llamaron la atención sobre la vasta multiplicidad de los elementos conductuales que pueden organizarse en grupos mediante enlaces hereditarios y de la experiencia.

Historial de la experiencia. Existe un reconocimiento creciente de la función que desempeñan las experiencias del individuo en el desarrollo de la inteligencia y la formación de factores de grupo (Anastasi, 1986b; Greeno, 1989). La experiencia no sólo influye en el nivel de desempeño en diferentes habilidades, sino también en la forma en que éste se organiza en distintos rasgos. Se ha descubierto que las diferencias en los patrones de factores se asocian con distintas culturas o subculturas, niveles socioeconómicos y programas académicos (véase Anastasi, 1970, 1983a, 1986b, 1994; Vernon, 1969), pero también son importantes las modificaciones que sufren estos patrones con el tiempo, incluyendo los cambios a largo plazo (que pueden reflejar el efecto acumulado de las experiencias cotidianas), los cambios a corto plazo que resultan de la práctica y otras experiencias de aprendizaje controladas experimentalmente (Baltes, Cornelius, Spiro, Nesselroade y Willis, 1980; Birren, Cunningham y Yamamoto, 1983; Fleishman, 1972; Fleishman y Mumford, 1989; Khan, 1970, 1972; Reinert, 1970). La investigación realizada con animales también ha arrojado evidencias interesantes sobre la producción experimental de factores mediante el control de las experiencias tempranas (Whimbey y Denenberg, 1966).

La composición factorial de la misma tarea objetiva suele diferir entre individuos con experiencias diversas, lo que puede obedecer al uso de métodos distintos para realizar la misma tarea; por ejemplo, quienes poseen habilidades verbales muy desarrolladas tenderán a utilizar mediadores verbales para resolver un problema espacial o mecánico, mientras que aquellos con experiencias predominantemente mecánicas se inclinarán a

mostrar un acercamiento perceptual o espacial al mismo problema. French (1965) ha aportado evidencias relevantes al encontrar que la composición factorial de los mismos instrumentos difería entre grupos de personas clasificadas de acuerdo con sus estilos peculiares de resolución de problemas. También hay evidencias interesantes en un estudio de C. H. Frederiksen (1969) sobre las estrategias cognoscitivas que utilizan los sujetos al memorizar palabras. En el curso del aprendizaje, los individuos pueden cambiar de estrategia y, en consecuencia, modificar la composición factorial de la tarea que realizan. Se ha demostrado que en el curso de la instrucción ocurren cambios similares en las aptitudes requeridas durante largos periodos (R. B. Burns, 1980).

Mecanismos de la formación de rasgos. Los conocidos conceptos de disposición para el aprendizaje y transferencia del entrenamiento ofrecen un mecanismo que da cuenta de la aparición de factores (Carroll, 1966; G. A. Ferguson, 1954, 1956; Simon, 1990; Whiteman, 1964). El establecimiento de la disposición para el aprendizaje permite al individuo aprender mejor cuando se le presenta un nuevo problema de la misma categoría. En los clásicos experimentos realizados por Harlow (1949, 1960) con monos, después de resolver problemas que requerían la diferenciación entre ciertas formas (como triángulos y círculos), el animal aprendía a discriminar *otras* formas con mayor rapidez que otros animales que carecían de esa experiencia. El mono había establecido una disposición para aprender a diferenciar formas, sabía lo que tenía que buscar al enfrentarse con un nuevo problema, es decir, había "aprendido cómo aprender" a resolver esta situación.

De modo similar, muchas de las habilidades desarrolladas durante la escolaridad formal, como la lectura y las operaciones aritméticas, pueden aplicarse luego a una gran variedad de situaciones de aprendizaje. De la misma manera, las técnicas eficaces y sistemáticas de resolución pueden aplicarse para resolver nuevos problemas. Las diferencias individuales en la adquisición de esas habilidades se verán reflejadas en la ejecución de un gran número de tareas diferentes, y, en el análisis factorial de éstas, las habilidades de aplicabilidad general surgirán como factores de grupo cuya amplitud estará determinada por la magnitud del efecto de transferencia o la variedad de tareas a las que puede aplicarse la habilidad.

Otra fuente importante de formación de rasgos es la proximidad u ocurrencia concurrente de las experiencias de aprendizaje; por ejemplo, es probable que las personas que provienen de culturas escolarizadas desarrollen un factor verbal-educativo, ya que éste se encuentra presente en todas las actividades académicas. El factor más limitado de la aptitud numérica puede resultar del hecho de que el mismo maestro enseña juntos todos los procesos aritméticos en el mismo salón de clases, de ahí que el niño desanimado, hostilizado o aburrido durante la clase de aritmética tenderá a rezagarse en el aprendizaje de *todos* esos procesos, mientras que quien es estimulado y gratificado mostrará un buen aprendizaje de todo lo que se le enseñe en el periodo y adquirirá aptitudes que le harán más fácil el aprendizaje numérico posterior.

Independientemente de cómo se desarrollen, los factores o las habilidades identificados mediante el análisis factorial son categorías descriptivas que reflejan las relaciones cambiantes del desempeño en diversas situaciones. Estos factores no son entidades estáticas, sino que son producto de la historia acumulada de experiencias del individuo. Es de esperar que en la medida en que varíen las correlaciones entre las

experiencias de individuos y grupos, existan entre ellos patrones diferentes. En la medida en que cambien las experiencias del individuo (sea por la educación formal, las funciones ocupacionales y otras actividades continuas) pueden diferenciarse nuevos rasgos o bien los que ya existen pueden surgir en compuestos más amplios.

Análisis factorial y análisis de las tareas cognoscitivas. Los más recientes procedimientos desarrollados por la psicología cognoscitiva sobre el procesamiento de información están haciendo contribuciones importantes no sólo a la comprensión de lo que miden los tests de inteligencia (véase el capítulo 5), sino también al entendimiento de cómo se forman y se desarrollan los factores. El análisis de protocolos (o registro), en el que el individuo debe "pensar en voz alta" mientras resuelve un problema o realiza una tarea intelectual, resulta una aproximación prometedora al análisis del pensamiento humano (Ericsson y Simon, 1993). Sin embargo, conforme prosigue la investigación de los procesos cognoscitivos crecen las evidencias respecto de la *especificidad de dominio de los procesos de pensamiento*. Excepto en los niveles más elementales, las habilidades de procesamiento tienden a ser específicas al contenido procesado, y se adquieren conforme se domina el conocimiento y se organiza para su posterior recuperación.

Por otro lado, los rasgos identificados con más frecuencia mediante el análisis factorial tienen que ver ante todo con dominios de contenido, como el verbal, numérico y espacial; e incluso los factores menos comunes que se definen por procesos, como la memoria asociativa, el razonamiento general o el pensamiento divergente, resultan ser específicos de un dominio cuando se evalúan por medio de pruebas especialmente desarrolladas. Independientemente de que un instrumento se derive del análisis factorial o del cognoscitivo, la distinción fundamental está entre el proceso y el dominio al que se aplica. El dominio relevante puede referirse al *contenido* que es procesado (por ejemplo, lingüístico, matemático o mecánico) o al *contexto* en el que se desempeña, según se identifique por las categorías cultural, social, geográfica, ocupacional u otras categorías ambientales.

El desarrollo de la psicología cognoscitiva, desde sus primeros esfuerzos por identificar los *procesos* de la inteligencia humana hasta su rápido avance en la exploración de la especificidad de dominio de todos los procesos cognoscitivos, se refleja en el impresionante crecimiento de las publicaciones que resaltan la especificidad de dominio.[10] También es digno de mención el que las pruebas elaboradas para evaluar los rasgos definidos por el contenido derivadas del análisis factorial hayan demostrado ser buenos predictores del desempeño cotidiano. Es probable que dichos factores puedan identificar indirectamente los agrupamientos de habilidades de procesamiento que son adecuadas para un determinado dominio de contenido. De este modo, quien obtiene una puntuación alta en una prueba de aptitud verbal no sólo puede destacar en el grado y la organización del conocimiento verbal, sino también en el uso de habilidades especializadas de procesamiento que se requieren en el dominio verbal.

Inteligencia general. La investigación sobre el análisis factorial de la inteligencia no sólo tiene que ver con lo que miden los tests de inteligencia, sino con la naturaleza

[10] Véase, por ejemplo, Greeno (1989), E. Hunt (1987), Schneider y Weinert (1990), Simon (1990), y Sternberg y Frensch (1991).

de la inteligencia, como quiera que se manifieste. Cuando se considera desde este punto de vista más amplio, la inteligencia humana comprende la combinación de habilidades cognoscitivas y de conocimiento que demanda, fomenta y recompensa el contexto de experiencias en el que funciona el individuo (Anastasi, 1986c). En este constructo generalizado, la especificidad de dominio es incluso más fundamental que en los procesos identificados con más claridad por el análisis de protocolo (o registro) o el análisis factorial. Muchos aspectos de la inteligencia se desarrollan mientras el individuo adquiere el conocimiento real y las habilidades de procesamiento en un dominio, como un contexto cultural o una ocupación específica.

Los primeros esfuerzos por identificar una única inteligencia de aplicación universal dieron lugar a los "tests de inteligencia" tradicionales, que llegaron a conocerse como medidas de inteligencia académica o de aptitud escolar por ser ésta la inteligencia que realmente medían. Se propuso entonces el término "inteligencia práctica" como un tipo distinto de inteligencia que no cubrían las pruebas habituales (véase Anastasi, 1986b; Neisser, 1976; Sternberg y Wagner, 1986). Sin embargo, la inteligencia práctica no es una sino una multiplicidad de inteligencias que pueden aplicarse en diferentes dominios prácticos (Lave, 1988; Rogoff y Lave, 1984). No existen dos tipos de inteligencia, académica y práctica; la inteligencia es un constructo multifacético.

Una debilidad del análisis factorial tradicional surge de la inadecuada atención que se le da a la selección de las variables que han de ser analizadas (Anastasi, 1988a). Obviamente, los factores identificados por el análisis se derivan de las correlaciones entre las variables elegidas. Casi todas las investigaciones de la inteligencia empezaron con un conjunto de puntuaciones sobre pruebas diseñadas para medir el rendimiento en la enseñanza escolar, de ahí que los factores que aparecieron representaran la inteligencia académica. En las culturas escolarizadas, estos factores no sólo demostraron ser buenos predictores del aprovechamiento escolar, sino que además mostraron correlaciones moderadamente altas con el desempeño en muchos trabajos para los cuales resulta importante la escolaridad. Sin embargo, si pretendemos evaluar la inteligencia en contextos más amplios necesitamos partir del desempeño de la gente en las actividades cotidianas que valora cada cultura. Aunque es sencillo atraer a los estudiantes como participantes en la investigación psicométrica, también es evidente ahora que se tiene acceso a adultos de diversas ocupaciones. La investigación puede organizarse de forma que además de beneficiar a los participantes produzca datos científicos bien controlados. A este respecto, cada vez son más los ejemplos de esta forma de investigación tanto en la industria como en los contextos agrícolas (por ejemplo, Fleishman y Reilly, 1992b; Lubinski y Dawis, 1992; Whyte, 1991).

Algunos de los procedimientos de la psicología cognoscitiva han puesto de relieve la especificidad de dominio de la inteligencia. La investigación de los procesos cognoscitivos realizada mediante el análisis de protocolo de grupos contrastados, como expertos y novatos en una actividad particular (digamos, ajedrecistas, mecanógrafas, programadores de cómputo, etc.), ha establecido una y otra vez que el nivel de desempeño y el uso de habilidades cognoscitivas de un individuo son estrictamente específicos del dominio particular. Una ventaja especial de este procedimiento de la psicología cognoscitiva es que la investigación se ha llevado a cabo mediante diversas actividades que tienen lugar en contextos reales, como es el caso de la memoria de los meseros para las

órdenes, el diagnóstico médico y las decisiones en las cortes judiciales (Chi *et al.*, 1988; Ericsson y Smith, 1991).

El concepto de especificidad de dominio ha tenido un efecto creciente en el estudio psicológico de diversos temas, como puede constatarse, por ejemplo, en las definiciones actuales de genio y talento. La identificación de niños sobredotados ha cambiado de determinado nivel de CI (Terman *et al.*, 1925) al de talento superior en cualquiera de los múltiples campos socialmente deseables (Csikszentmihalyi, Rathunde y Whalen, 1993; Feldman y Bratton, 1972; Horowitz y O'Brien, 1985; Subotnik y Arnold, 1994). De igual modo, los primeros esfuerzos por diseñar pruebas de procesos generales de creatividad o pensamiento divergente encontraron una clara especificidad de dominio (Baer, 1993; Runco, 1991, 1994; Subotnik y Arnold, 1994). Actualmente se observa cierto progreso en la integración del dominio en las pruebas de creatividad, lo que permite suponer que la unión de la psicología cognoscitiva con el análisis factorial no sólo ha enriquecido la comprensión del funcionamiento intelectual, sino que también ha permitido que ambas investigaciones se acerquen más a la vida cotidiana.

Capítulo Doce

Consecuencias psicológicas en la evaluación de la habilidad

Una consecuencia inevitable de la expansión y la mayor complejidad de cualquier empresa científica es la especialización de los intereses y las funciones de quienes la practican, lo que resulta evidente en la relación que guarda la evaluación psicológica con las tendencias de la psicología contemporánea (Anastasi, 1967, 1991). Los psicómetras han establecido técnicas de muy elevada calidad para la elaboración de pruebas, pero no han tenido el cuidado de cerciorarse de que los usuarios cuenten con la información psicológica indispensable para hacer un uso adecuado de sus instrumentos, lo que ha ocasionado que persistan interpretaciones anticuadas del desempeño que no hacen referencia a los resultados de la investigación conductual pertinente. Buena parte del descontento público con la evaluación psicológica —que empezó en la década de los cincuenta, alcanzó niveles escandalosos en los setenta y continua aún en muchos contextos— se explica por este aislamiento parcial que tiene respecto a otros campos de la psicología (con los consecuentes abusos y malas interpretaciones de las pruebas). El tema que veremos en este capítulo ilustra la contribución de los descubrimientos de la investigación psicológica al uso eficaz de las pruebas de habilidad y a corregir las creencias erróneas sobre el "CI" y otras calificaciones similares.

ESTUDIOS LONGITUDINALES DE LA INTELIGENCIA INFANTIL

Los estudios longitudinales de los mismos individuos a lo largo de periodos prolongados constituyen una forma importante para aproximarse a la comprensión del constructo de "inteligencia". Aunque es factible considerar que estas investigaciones contribuyen a la validación predictiva a largo plazo de instrumentos específicos, tie-

nen mayores implicaciones para la naturaleza de la inteligencia y el significado de las puntuaciones de las pruebas que pretenden medirla. Cuando se suponía que la inteligencia era en gran medida la expresión de un potencial hereditario, se esperaba que el CI de cada individuo permaneciera inmutable durante su vida. Cualquier variación observada en una nueva evaluación se atribuía a un defecto del instrumento de medición, ya fuera por un problema de confiabilidad o por la mala selección de las funciones evaluadas. Sin embargo, al aumentar las investigaciones sobre la naturaleza de la inteligencia llegó a reconocerse que ésta es compleja y dinámica. En las siguientes secciones examinaremos los descubrimientos más frecuentes de los estudios longitudinales de la inteligencia y buscaremos las condiciones que contribuyen a la estabilidad y a la inestabilidad de las habilidades examinadas.

Estabilidad de la ejecución en los tests de inteligencia. Se cuenta con una cantidad importante de datos que muestra que la ejecución en los tests de inteligencia permanece considerablemente estable desde la escuela elemental hasta la secundaria y la universidad (véase Anastasi, 1958, pp. 232–238; Bornstein y Krasnegor, 1989; McCall, Appelbaum y Hogarty, 1973). Por ejemplo, en un estudio sueco de una población relativamente no seleccionada, Husén (1951) encontró una correlación de .72 entre las calificaciones obtenidas por 613 estudiantes varones cuando cursaban el tercer grado de secundaria y las calificaciones que obtuvieron 10 años más tarde durante la inducción al servicio militar. En un estudio sueco posterior, Härnqvist (1968) descubrió una correlación de .78 entre las pruebas aplicadas a los 13 y a los 18 años de edad a más de 4 500 varones. Incluso las pruebas aplicadas a preescolares muestran correlaciones notablemente elevadas con el retest posterior. En un estudio longitudinal conducido en el Instituto de Investigación Fels con 140 niños (Sontag, Baker y Nelson, 1958), se obtuvo una correlación de .83 entre las calificaciones Stanford-Binet obtenidas a los tres y a los cuatro años de edad; aunque la correlación de la prueba aplicada a los tres años disminuía al aumentar el intervalo entre el retest, a los 12 años todavía alcanzaba a ser de .46. De especial relevancia para el Stanford-Binet resulta el seguimiento realizado por Bradway, Thompson y Cravens (1958) en niños que ya habían sido evaluados entre los dos y los cinco años y medio como parte de la muestra de estandarización del Stanford-Binet de 1937. Los CI originales correlacionaron .65 en el retest aplicado a los 10 años y .59 en el aplicado a los 25. La correlación entre el retest aplicado a los 10 años (edad promedio = 14 años) y el aplicado 25 años después (edad promedio = 29 años) fue de .85.

Como era de esperarse, las correlaciones en el retest aumentan al disminuir el intervalo entre las pruebas; pero si se mantiene constante, las correlaciones en el retest tienden a ser más altas en los niños de mayor edad. Los efectos de la edad y el intervalo entre las evaluaciones sobre las correlaciones observadas en el retest son notablemente regulares y tienden a ser muy predecibles (R. L. Thorndike, 1933, 1940), lo que puede explicarse por la *naturaleza acumulativa del desarrollo intelectual*. En cualquier nivel de edad, el conocimiento y las habilidades intelectuales del individuo incluyen, además de las nuevas adquisiciones, los conocimientos y las habilidades aprendidos antes. Incluso si los incrementos anuales no estuvieran relacionados, puede esperarse que aumente la consistencia del nivel de ejecución sencillamente porque al aumentar la edad las adquisiciones previas constituyen una proporción creciente del total de habilidades y conocimientos.

Aunque el traslape de habilidades y conocimientos en las edades sucesivas puede dar cuenta de la mayor estabilidad de las puntuaciones obtenidas en las pruebas por el individuo en desarrollo, otras dos condiciones merecen ser consideradas. La primera es la *estabilidad ambiental* que caracteriza a los años de desarrollo de casi todas las personas, de ahí que cualquier ventaja o desventaja que alguien haya sufrido en una etapa suele persistir en el intervalo entre el retest. Mientras crecen, los niños tienden a permanecer en el mismo núcleo familiar, a mantener el nivel socioeconómico y el mismo medio cultural ya que, por lo general, no mudan al azar de un ambiente de estímulos intelectuales a uno empobrecido.

Al mismo tiempo, observemos que los ambientes psicológicos de los hermanos distan mucho de ser idénticos: En efecto, cada vez son más los datos que muestran que las historias ambientales de los hermanos criados juntos pueden diferir de muchas maneras (Boer y Dunn, 1992; Dunn y Plomin, 1990; Hetherington, Reiss y Plomin, 1993). Tener un hermano mayor o menor es una experiencia psicológica significativamente diferente al hecho de ser hijo único. La naturaleza de las prácticas de crianza infantil, así como la propia respuesta de los padres a la conducta del niño, también pueden variar entre hermanos nacidos en tiempos diferentes. Más aún, los acontecimientos importantes de la vida familiar (como el divorcio, los cambios en el ingreso familiar o el traslado de un ambiente rural a uno urbano) afectan de manera diferente a los niños, dado que se encuentran en etapas distintas de desarrollo. Todas estas posibles diferencias deben agregarse a las experiencias acumulativas que los hermanos viven fuera de la familia; por consecuencia, aunque la continuidad familiar favorezca a la estabilidad en la ejecución que el niño muestra en una prueba, no necesariamente contribuye al parecido entre hermanos.

La segunda condición que promueve la estabilidad general de la ejecución del individuo en los tests de inteligencia tiene que ver con la función de las *habilidades que constituyen un requisito para el aprendizaje* posterior. El individuo no sólo debe conservar lo que ha aprendido, sino que en buena medida le proporciona las herramientas para el aprendizaje posterior; por lo tanto, entre más progrese un niño en cualquier momento en la adquisición de conocimientos y las habilidades intelectuales, mayor será la probabilidad de que se beneficie de las experiencias posteriores. En la educación, el concepto de preparación es una expresión de este principio general. La naturaleza secuencial del aprendizaje está implicada en la aproximación piagetiana al desarrollo mental y en varios programas de instrucción individualizada.

Las aplicaciones de este principio constituyen el fundamento de proyectos como el Head Start y otros programas de educación compensatoria para niños que sufren desventajas educativas (Stanley, 1972, 1973; Zigler y Valentine, 1980). En la medida en que los niños que provienen de medios socioeconómicos muy bajos carecen de los requisitos básicos para el aprendizaje escolar eficaz, su rendimiento académico tenderá a rezagarse conforme vayan avanzando en los grados escolares. Debe agregarse que los requisitos de aprendizaje no sólo abarcan habilidades intelectuales como la adquisición del lenguaje y de conceptos cuantitativos, sino también actitudes, intereses, motivación, estilos de resolución de problemas, reacciones ante la frustración, autoconcepto y otras características de personalidad. El propósito de los programas de educación compensatoria es proporcionar a esos niños los requisitos de aprendizaje que les permitan beneficiarse de la escolaridad subsecuente; por supuesto, tales

programas esperan interrumpir la "estabilidad" del CI, que de otra manera habría permanecido bajo.

Inestabilidad de la ejecución en los tests de inteligencia. Los estudios sobre la correlación de las puntuaciones de las pruebas proporcionan datos numéricos que pueden utilizarse para hacer predicciones de grupo. Por las razones ya mencionadas, las puntuaciones tienden a permanecer estables en el sentido numérico, pero a nivel individual es posible encontrar considerables aumentos y disminuciones que obedezcan a cambios ambientales importantes en la vida del niño; por ejemplo, su desarrollo intelectual puede verse afectado por la ocurrencia de factores como cambios drásticos en la estructura familiar o en las condiciones del hogar, adopción, enfermedades graves o prolongadas y la participación en programas terapéuticos o de remedio. Incluso los niños que permanecen en el mismo ambiente pueden mostrar aumentos o disminuciones notables en las puntuaciones que obtienen al aplicarles nuevamente una prueba. Como es obvio, esos cambios significan que el niño se está desarrollando a un ritmo más rápido o más lento que el de la población normativa con la que se estandarizó la prueba. En general, conforme crecen los niños que provienen de ambientes con desventajas educativas sus puntuaciones tienden a disminuir, mientras que los que se encuentran en ambientes favorables muestran aumentos. En nuestros días, la investigación está prestando mayor atención a las características específicas tanto de los ambientes como de los niños (véase por ejemplo, Carroll, 1993, pp. 669–674; Detterman y Sternberg, 1982).

El Estudio de Orientación de California (*California Guidance Study*; Honzik, Macfarlane y Allen, 1948) proporcionó una gran cantidad de información sobre la magnitud de los cambios individuales. Al analizar los datos del retest de 222 casos se encontraron cambios individuales en el CI de hasta 50 puntos. Luego del periodo de seis a 18 años, cuando las correlaciones del retest son generalmente elevadas, 59 por ciento de los niños mostró cambios de 15 o más puntos de CI, 37 por ciento en 20 puntos o más, y nueve por ciento en 30 o más. La mayor parte de esos cambios no fue de naturaleza aleatoria o errática; por el contrario, los niños mostraron tendencias constantes hacia arriba o hacia abajo luego de varios años consecutivos, y a su vez las modificaciones se relacionaron con las características del ambiente. En el Estudio de Orientación de California, la investigación detallada de las condiciones del hogar y de las relaciones entre padres e hijos demostró que los aumentos o las disminuciones en el CI se asociaban con el medio cultural y el clima emocional del desarrollo del niño. Un seguimiento conducido cuando los participantes alcanzaron la edad de 30 años todavía encontró correlaciones significativas entre las puntuaciones obtenidas en las pruebas y el medio familiar evaluado los 21 meses de edad (Honzik, 1967). La preocupación de los padres por el aprovechamiento educativo del hijo surgió como correlato importante de la ejecución subsecuente en la prueba, al igual que otras variables que reflejan el interés de los progenitores por el bienestar general del niño.

Algunos investigadores se concentraron en concreto en las características de personalidad asociadas con el incremento y el decremento intelectual. En el Instituto de Investigación Fels, se incluyó a 140 niños en un estudio longitudinal que se extendió desde la primera infancia hasta después de la adolescencia (Kagan y Freeman, 1963; Kagan, Sontag, Baker y Nelson, 1958; Sontag *et al.*, 1958). Se compararon los niños que mostraban las mayores ganancias y los que presentaban las mayores pérdidas en el

CI entre las edades de cuatro y medio y seis años en una variedad de medidas ambientales y de personalidad; lo mismo se hizo con los que mostraban los mayores cambios en el CI entre los seis y los 10 años. Durante los años preescolares, la dependencia emocional de los padres fue la principal condición asociada con las pérdidas de CI; en los años escolares, las ganancias en el CI se asociaron principalmente con la mayor motivación de logro, los esfuerzos de superación y la curiosidad por la naturaleza. Se obtuvieron también datos interesantes sobre la función de la actitud de los padres y sus prácticas de crianza en el desarrollo de esos rasgos.

Un análisis posterior de la misma muestra, que se extendió hasta los 17 años, se concentró principalmente en los patrones de cambio de puntuación observados con el paso del tiempo (McCall *et al.*, 1973). Se comparó a los niños que mostraban patrones diferentes en relación con las prácticas de crianza infantil evaluadas durante las visitas periódicas al hogar. Un descubrimiento habitual fue que los padres de los niños cuyas puntuaciones mostraban una tendencia a aumentar durante los años preescolares presentaban "una atmósfera estimulante y remuneradora, pero con cierta estructura y reglas establecidas" (McCall *et al.*, 1973, p. 54). Una condición importante asociada con el aumento de las puntuaciones se describe como un intento de aceleración, o el grado en el que "el padre deliberadamente da entrenamiento al niño en diversas habilidades mentales y motoras que aún no eran esenciales" (p. 52).

La investigación sobre los factores asociados con los aumentos y las disminuciones de las puntuaciones de los tests de inteligencia ha permitido identificar las condiciones que determinan el desarrollo intelectual en general, a la vez que postula la posibilidad de mejorar la predicción del estado intelectual subsecuente, combinando las medidas de las características emocionales y motivacionales del individuo con las medidas de su ambiente y las puntuaciones iniciales de las pruebas. Desde otro punto de vista, los descubrimientos de estas investigaciones indican los programas de intervención que pueden modificar de manera eficaz el curso del desarrollo intelectual en la dirección deseada.

LA INTELIGENCIA EN LA NIÑEZ TEMPRANA

La evaluación de la inteligencia en los dos extremos del rango de edad presenta problemas teóricos y de interpretación especiales. Uno de ellos tiene que ver con las funciones que deben evaluarse, es decir, determinar qué constituye la inteligencia para el infante y el preescolar, por un lado, y qué constituye la inteligencia para el viejo, por el otro. El segundo problema no es del todo independiente del primero: a diferencia del escolar, infantes y preescolares no han sido expuestos a la serie estandarizada de experiencias del programa escolar. Al desarrollar pruebas para los niveles de educación elemental, secundaria y universitaria, los autores disponen de un caudal de experiencias comunes del que pueden extraer los reactivos de sus pruebas. Por otra parte, y a pesar de ciertas uniformidades culturales en las prácticas de crianza infantil, las experiencias del niño antes del ingreso a la escuela son mucho menos homogéneas, lo que hace mucho más difícil tanto la elaboración de las pruebas como la interpretación de sus resultados. Hasta cierto punto, se enfrenta el mismo problema al evaluar a los viejos, cuya escolaridad concluyó muchos años atrás y que desde entonces han vivido experiencias muy diferentes. En esta sección y la próxima examinaremos algunas de las implicaciones de estos problemas en la evaluación de niños y adultos, respectivamente.

Validez predictiva de las pruebas para infantes y preescolares. La conclusión que surge de los estudios longitudinales es que las pruebas para preescolares (en especial cuando se aplican después de los dos años de edad) tienen una validez moderada para predecir la ejecución posterior en tests de inteligencia, y que las pruebas para infantes prácticamente no tienen validez alguna (Bayley, 1970; Lewis, 1973; McCall, Hogarty y Hurlburt, 1972). Al combinar los resultados de ocho estudios, McCall y sus colaboradores (1972) calcularon las correlaciones promedio entre las pruebas aplicadas en los primeros 30 meses de vida y el CI obtenido entre los tres y los 18 años. De sus datos resultan evidentes varias tendencias. Primera, las pruebas aplicadas en el primer año de vida tienen poco o ningún valor predictivo a largo plazo. Segunda, las pruebas para infantes muestran cierta validez para predecir el CI en edades preescolares (de tres a cuatro años), pero la correlación muestra un brusco descenso después de este punto, luego de que el niño alcanza la edad escolar. Tercera, después de los 18 meses de edad la validez predictiva es moderada y estable, ubicada en general entre .40 y .50. Cuando se hacen predicciones a partir de esas edades, las correlaciones parecen ser del mismo orden de magnitud, independientemente de lo que en el retest ocurra en cualquier momento de los tres a los 18 años.

Es necesario evaluar la falta de validez predictiva de las pruebas para infantes en relación con otros descubrimientos. Primero, las predicciones pueden mejorar si se consideran las tendencias del desarrollo con evaluaciones repetidas. Segundo, algunos investigadores han encontrado que las pruebas para infantes tienen una mayor validez predictiva en poblaciones clínicas anormales que en poblaciones normales; por ejemplo, se han informado coeficientes de validez de .60 y .70 para niños con un CI inicial inferior a 80, así como en grupos que presentan (o en los que se sospecha) anormalidades neurológicas (Ireton, Thwing y Gravem, 1970; Knobloch y Pasamanick, 1963, 1966; Werner, Honzik y Smith, 1968). Las pruebas para infantes, pues, parecen ser más útiles como auxiliares en el diagnóstico de problemas del desarrollo causados por patologías orgánicas o que tienen otro origen hereditario o ambiental.

En ausencia de alguna patología, el desarrollo posterior del niño está determinado sobre todo por el ambiente en el que crece, que no se espera que la prueba pueda predecir. En efecto, la educación de los padres y otras características más específicas del entorno familiar son mejores predictores del nivel intelectual subsecuente que las puntuaciones de las pruebas infantiles; después de los 18 meses, la predicción mejora considerablemente si las puntuaciones se combinan con indicadores de la posición socioeconómica de la familia (Bayley, 1955; McCall *et al.*, 1972; Pinneau, 1961; Werner *et al.*, 1968). Se ha propuesto que las diferencias individuales en la infancia pueden ser relativamente menores y transitorias, ya que el desarrollo normal es en esencia común a la especie en esta etapa temprana (R. B. McCall, 1981). En los años posteriores, las diferencias individuales aumentan, se hacen más estables y dan lugar a mayores correlaciones con los factores genéticos y ambientales (Plomin, De Fries y Fulker, 1988). Pese a ello, debe advertirse que en la década de los noventa ha crecido la investigación sobre el valor predictivo de la conducta cognoscitiva de los infantes, con resultados prometedores (Colombo, 1993).

Naturaleza de la inteligencia en la niñez temprana. Si consideramos la investigación sobre la naturaleza de la inteligencia en la niñez temprana —cuyos descubrimientos no respaldan la idea de que el desarrollo de las capacidades intelectuales en

la infancia sea unitario y constante (Lewis, 1973, 1976; McCall *et al.*, 1972)— podremos entender más plenamente la validez de los tests de inteligencia infantil y el significado de las medidas de la ejecución temprana. Las correlaciones son mínimas incluso con intervalos de tres meses, y las correlaciones con la ejecución en las mismas escalas u otras a los dos o más años de edad suelen ser insignificantes. Estos resultados se han obtenido lo mismo con instrumentos estandarizados —como las Escalas de Bayley sobre el Desarrollo Infantil— que con escalas ordinales, como las piagetianas (Gottfried y Brody, 1975; King y Seegmiller, 1973; Lewis, 1976; Lewis y McGurk, 1972).

Varios investigadores han llegado a la conclusión de que si bien los tests de inteligencia infantil carecen de validez predictiva, son indicadores válidos de las habilidades cognoscitivas que el niño presenta en ese momento (Bayley, 1970; Stott y Ball, 1965; Thomas, 1970). Según este punto de vista, el hecho de que la naturaleza y la composición de la inteligencia cambian con la edad es un factor que explica lo insignificante de las correlaciones entre la pruebas infantiles y la ejecución posterior; la inteligencia en la infancia es cualitativamente diferente de la que se observa en la edad escolar y consiste en una combinación distinta de habilidades.

McCall y sus colaboradores exploraron en una serie de investigaciones la naturaleza cambiante de la inteligencia infantil a intervalos de seis meses durante los dos primeros años de vida (R. B. McCall, 1976; McCall, Eichorn y Hogarty, 1977; McCall *et al.*, 1972). Estos especialistas buscaron los precursores del desarrollo posterior de la conducta infantil mediante el análisis estadístico de las correlaciones de diferentes habilidades en cada nivel de seis meses y de las correlaciones de las mismas habilidades y de otras distintas en diferentes edades. Una de las conclusiones a que llegó la investigación fue que la conducta predominante en diferentes edades muestra cambios cualitativos que representan transiciones ordenadas y razonables. Cuando las respuestas de los infantes a los reactivos de Gesell se sometieron por separado a un análisis factorial en niveles sucesivos de seis meses, las calificaciones del primer factor en cada nivel mostraron una correlación significativa entre las edades; pese a ello, la composición conductual de estos primeros factores variaba con la edad, es decir, las manifestaciones específicas de la competencia mental diferían según los niveles de edad, aunque la competencia en una edad predecía la de edades posteriores cuando se evaluaban por medio de las conductas apropiadas para la edad.

Se introdujo el concepto de *transformaciones del desarrollo* para describir los cambios en las manifestaciones apropiadas para la edad en la competencia intelectual. Las evidencias posteriores de los cambios cualitativos en la competencia conductual se encuentran en la investigación de Yarrow y sus colaboradores sobre la forma en que el infante domina su entorno (Messer *et al.*, 1986; Morgan y Harmon, 1984; Yarrow *et al.*, 1983; Yarrow *et al.*, 1984; Yarrow y Messer, 1983). Los resultados demostraron una progresión temporal tanto en las tareas que daban lugar a esta conducta como en la clase específica de conducta provocada: la inspección visual, la manipulación y la persistencia en la solución de problemas. Al principio, el infante descubre que puede modificar el medio, por ejemplo al tirar un cubo para verlo caer y escucharlo golpear el piso o al agitar una campana para oír su sonido. Después, el dominio del ambiente se manifiesta en actividades más complejas dirigidas a una meta, como cuando utiliza desviaciones o relaciones entre medios y fines para obtener un juguete. Al identificar la conducta específica apropiada para la edad es posible investigar de mejor manera tanto la validez de constructo como la validez predictiva de la evaluación de la inteligencia en la niñez temprana. También es

importante considerar la función que cumple el contenido del conocimiento en los procesos intelectuales y las estrategias cognoscitivas (Reese, 1987), un hecho que cada vez obtiene mayor reconocimiento en la investigación de la psicología cognoscitiva.

Implicaciones de los programas de intervención. La eficacia de los diversos programas de intervención de la época del Head Start depende de la calidad de cada uno (R. C. Collins, 1993; Haskins, 1989; Zigler y Muenchow, 1992; Zigler y Styfco, 1993). Sus procedimientos y resultados, que estaban destinados a mejorar la preparación académica de los niños que provenían de medios desfavorecidos, mostraban grandes variaciones. En su mayor parte se trataba de proyectos iniciados apresuradamente, con una planeación inadecuada tanto para su implantación como para su evaluación, por lo que sólo algunos pudieron demostrar mejorías sustanciales en la ejecución de los niños, aunque por lo general eran progresos limitados y de corta vida. En contraste con la mayor parte de los programas, que buscaban "elevar el CI" mediante procedimientos vagamente definidos, unos cuantos proyectos de alta calidad precisaron de manera específica las habilidades intelectuales que pretendían mejorar y seleccionaron los procedimientos adecuados de entrenamiento. En estos casos, seguimientos cuidadosamente conducidos mostraron progresos significativos y duraderos en las habilidades relevantes. También prestaron atención al contexto más amplio en el que se introducía el programa, que podía incluir la necesidad de servicios sociales y de salud en el hogar. La participación de los padres demostró ser particularmente valiosa para complementar en casa el programa preescolar y asegurar su continuación cuando el proyecto oficial hubiese terminado (Jaynes y Wlodkowski, 1990).

También merecen atención los programas de seguimiento para evaluar la naturaleza y duración de la intervención, ya que la evaluación de la eficacia de dichos proyectos requiere considerable sofisticación metodológica (Collins y Horn, 1991; Willett y Sayer, 1994). Además de las cuestiones del diseño experimental, los artefactos estadísticos asociados con las propiedades psicométricas de los instrumentos de evaluación pueden dar lugar a falsos resultados positivos o negativos (Béjar, 1980). Las diferencias en la dificultad o la capacidad discriminativa de los reactivos entre los grupos experimental y de control, o bien entre la ejecución en el pretest y el postest del mismo grupo, pueden conducir a conclusiones incorrectas sobre el éxito o el fracaso del programa. Es posible evitar algunos de estos problemas con el uso de las pruebas desarrolladas y calificadas sobre la base de la teoría de respuesta al ítem (TRI, capítulo 7) y ajustadas individualmente a los examinados (CAT, capítulo 11). En los años recientes se ha observado un interés renovado por establecer programas de intervención bien diseñados y ejecutados con cuidado que tomen como modelo algunos de los programas anteriores más eficaces (R. C. Collins, 1993; Consortium, 1983; Haskins, 1989; Whimbey, 1990; Zigler y Styfco, 1993). Más aún, el desarrollo de esos nuevos programas puede beneficiarse de la base de datos formada a partir de la investigación de la inteligencia infantil (véase, por ejemplo, Horowitz y O'Brien, 1989). Un proyecto a largo plazo en especial prometedor se concentra en el efecto de la conducta de los padres hacia sus hijos de uno y dos años en el desempeño intelectual que posteriormente muestren los niños (Hart y Risley, 1995). Los primeros datos ya presentan un fuerte apoyo a la existencia de una relación estrecha entre la naturaleza y el grado de contactos paternos y el desarrollo intelectual de los niños.

PROBLEMAS EN LA EVALUACIÓN DE LA INTELIGENCIA DE LOS ADULTOS

Decrementos con la edad. Un rasgo distintivo de las escalas de Wechsler para medir la inteligencia de los adultos (capítulo 8) era el uso de una norma declinante para calcular los CI de desviación. Las puntuaciones crudas de los subtests del WAIS (y del WAIS-R) primero se transformaban en calificaciones estándares con una media de 10 y una *DE* igual a 3. Las calificaciones escaladas se expresan en términos de los grupos fijos de referencia formados por 500 personas de entre 20 y 34 años incluidas en las muestras de estandarización. Con la suma de las calificaciones escaladas de los 11 subtests se encuentra el CI de desviación en la tabla apropiada de edad; sin embargo, si examinamos directamente las sumas, podemos comparar la ejecución de diferentes grupos de edad en términos de una sola escala continua. La figura 12.1 muestra las medias de esas calificaciones escaladas totales para los niveles de edad incluidos en la muestra nacional de estandarización.

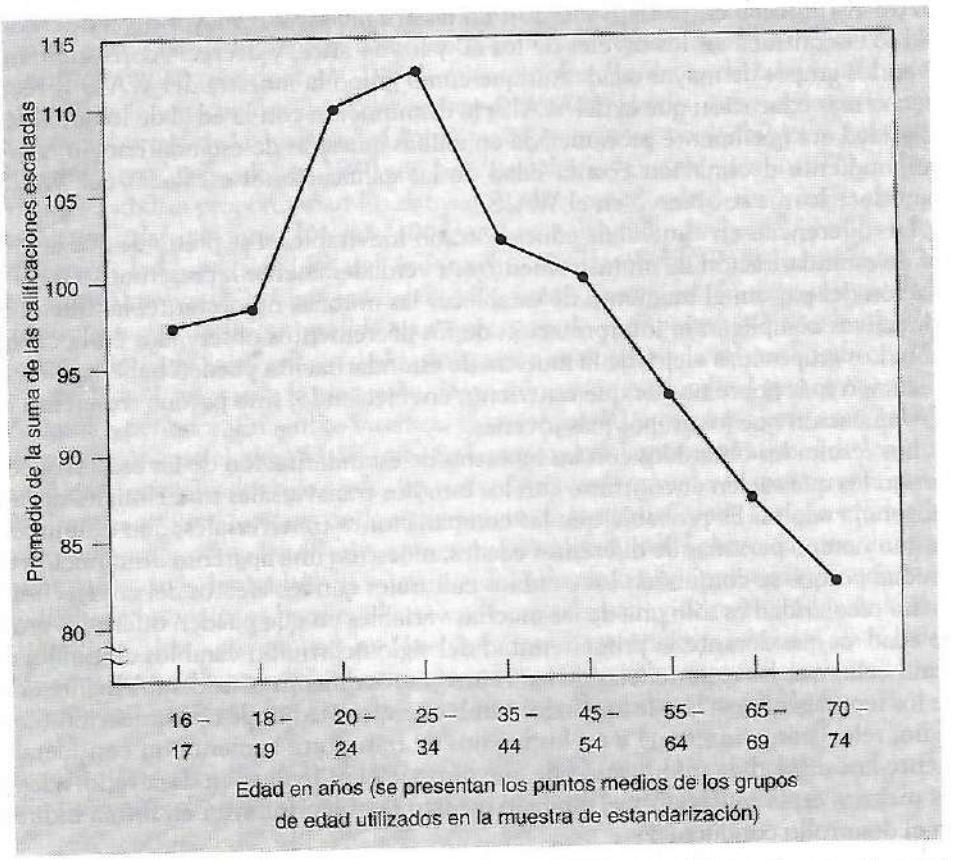


Figura 12.1. Disminución de las calificaciones escaladas de la escala completa del WAIS-R a lo largo de los conjuntos de edad.

(Datos tomados de Wechsler, 1981, p. 26.)

Como puede observarse en la figura 12.1, las calificaciones alcanzan un pico entre los 20 y los 34 años de edad para luego declinar de manera estable en todos los grupos de mayor edad. El CI de desviación se encuentra al referir la calificación escalada total del individuo a la norma de su grupo de edad. De esta manera, si el examinado muestra con la edad la misma declinación en la ejecución que la muestra normativa, su CI permanecerá constante. Según la lógica de Wechsler para usar este procedimiento, era "normal" que la capacidad de la persona examinada declinara con la edad después de los 30.

Sin embargo, al interpretar la declinación con la edad que se ilustra en la figura 12.1 debemos considerar un aspecto esencial de las muestras empleadas en la estandarización del instrumento. Como toda muestra de estandarización es una muestra *normativa*, debe reflejar las características de la población en cada nivel de edad (Anastasi, 1956), de lo que se deduce que cuando el nivel educativo de la población general se ha elevado durante varias décadas, los grupos de mayor edad habrán recibido menos educación que los más jóvenes. Esta diferencia educativa se refleja claramente en las muestras de estandarización del WAIS (probada entre 1953 y 1954) y el WAIS-R (probada entre 1976 y 1980). En las dos muestras el mayor nivel de escolaridad se encontraba en los niveles de los 20 y los 34 años, y decrecía consistentemente en los grupos de mayor edad. Aunque como grupo la muestra del WAIS-R recibió mucha más educación que la del WAIS, la disminución con la edad de los años de escolaridad era igualmente pronunciada en ambas muestras de estandarización, y la correspondiente declinación con la edad de las calificaciones escaladas del WAIS-R reproduce lo que se observó en el WAIS.

Las diferencias en el nivel de educación son inevitables si se pretende que la muestra de estandarización de un instrumento sea verdaderamente representativa de la población del país en el momento de establecer las normas; no obstante, las diferencias educativas complican la interpretación de los decrementos observados en la calificación. Los grupos más viejos de la muestra de estandarización pueden haber tenido una ejecución más pobre no por que estuvieran envejeciendo, sino porque recibieron menos educación que los grupos más jóvenes.

Los resultados obtenidos con las muestras de estandarización de las escalas Wechsler son los que suelen encontrarse con los estudios transversales tradicionales de la inteligencia adulta. Es probable que las comparaciones transversales, que examinan al mismo tiempo personas de diferentes edades, muestren una aparente disminución con la edad porque se confunden los cambios culturales con los efectos del envejecimiento. La escolaridad es sólo una de las muchas variables en que pueden diferir los grupos de edad ya que durante la primera mitad del siglo ocurrieron cambios culturales que han hecho que las experiencias de los veinteañeros difieran considerablemente de las de los septuagenarios. Desde luego, los cambios en los medios de comunicación (como radio, televisión e internet) y en los medios de transporte aumentaron considerablemente la cantidad de información de que dispone el individuo en desarrollo; además, las mejoras en la nutrición y el cuidado médico también influyen en forma indirecta en el desarrollo conductual.

En general, los estudios longitudinales, basados en la nueva evaluación de las mismas personas a lo largo de periodos que fluctúan entre los cinco y los 40 años, han revelado la tendencia opuesta, es decir, las calificaciones tienden a mejorar con la edad. Algunas de las investigaciones longitudinales han sido realizadas con grupos de nivel intelectual

superior, como los graduados universitarios o individuos elegidos por su CI elevado (Bayley y Oden, 1955; R. B. Burns, 1966; D. P. Campbell, 1965; Nisbet, 1957; Owens, 1953, 1966), por lo que algunos autores han argumentado que estos descubrimientos no se aplican a la población en general y deben restringirse a las personas con mayor nivel intelectual o educativo. Sin embargo, se han obtenido resultados similares en otros estudios longitudinales con muestras intermedias (Charles y James, 1964; Eisdorfer, 1963; Tuddenham, Blumenkrantz y Wilkin, 1968) y con adultos con retardo mental no recluidos (Baller, Charles y Miller, 1967; Bell y Zubek, 1960; Charles, 1953).

Ni los estudios transversales ni los longitudinales pueden explicar por sí mismos de manera concluyente los cambios observados con la edad. Por otro lado, en los estudios transversales las diferencias con la edad en el nivel educativo pueden dar lugar a una falsa disminución de la ejecución en la prueba relacionada con la edad. Además, conforme la gente envejece se ve expuesta a cambios culturales que pueden mejorar su desempeño en los tests de inteligencia. A este respecto, se han publicado[1] varios análisis excelentes de los problemas metodológicos que son inherentes a cada procedimiento junto con los diseños experimentales requeridos. Básicamente, se necesita la combinación de varios procedimientos, como en el *diseño de cruzamiento secuencial* (K. W. Schaie, 1965, 1994; Shock *et al.*, 1984), que combina los datos de las evaluaciones transversales y longitudinales en *comparaciones de series temporales*, que requieren la evaluación de cohortes[2] de la misma edad en momentos distintos; por ejemplo, se compara a los veinteañeros evaluados en 1940 con los veinteañeros examinados en 1970.

Algunos estudios proporcionan los datos que permiten al menos un análisis parcial de los factores participantes. Owens (1966) volvió a examinar luego de 40 años a los estudiantes de la Universidad Estatal de Iowa y D. P. Campbell (1965) lo hizo a los 25 años con los alumnos de la Universidad de Minnesota. Pero también examinaron a los *estudiantes actuales* de las universidades respectivas, lo que permitió hacer múltiples comparaciones entre la ejecución de los dos grupos evaluados a la misma edad con distancias de 25 a 40 años y de la ejecución de un solo grupo evaluado al principio y al final de los dos periodos. En ambos estudios, el grupo inicial mostró mejoras en relación con su ejecución inicial, pero su desempeño fue aproximadamente similar al de los grupos más jóvenes evaluados más tarde. Estos descubrimientos sugieren que son los cambios culturales y otros factores de la experiencia, más que la edad, los que producen tanto los aumentos como las disminuciones en las calificaciones obtenidas con diseños experimentales más limitados. El aumento del interés en la investigación sobre el aprendizaje de los viejos se hace evidente en la exhaustiva revisión realizada por Kausler (1994), cuyo libro ciertamente ayuda a despejar los estereotipos sobre los efectos del envejecimiento en el aprendizaje, y aunque se basa en descubrimientos científicos bien documentados, también ofrece muchos vínculos con la experiencia cotidiana.

El Estudio Longitudinal de Seattle (ELS). El Estudio Longitudinal de Seattle (*Seatle Longitudinal Study, SLS*) es un programa de investigación a largo plazo que

[1] Véase, por ejemplo, Baltes (1968), Botwinik (1984, capítulos 20 y 21), Buss (1973), Nesselroade y Reese (1973), Nesselroade y Von Eye (1985), K. W. Schaie (1973, 1988a) y Schaie y Hertzog (1986).

[2] Como se usa en este contexto, la cohorte es un grupo de personas de la misma edad, es decir, que nacieron en el mismo año o en otro periodo especificado.

además de estar particularmente bien planeado utiliza el diseño de cruzamiento secuencial (K. W. Schaie, 1994; Schaie y Hertzog, 1986). A partir de 1956, los investigadores aplicaron una batería de pruebas de habilidad[3] a una muestra aleatoria estratificada de 500 personas, tomadas de una población de aproximadamente 18 000 miembros de un plan de seguro de gastos médicos. La población se consideró bastante representativa de las cifras del censo de una importante área metropolitana: incluía a 25 hombres y 25 mujeres en cada intervalo de cinco años de los 21 a los 70. En las etapas sucesivas del estudio, se aplicaron las pruebas a lo largo de seis ciclos (de 1956 a 1991), en cada uno de los cuales no sólo se incluía a los supervivientes de la muestra original, sino que también se incluían nuevas muestras.

El diseño experimental básico del estudio incluyó el retest longitudinal repetido de las mismas personas, comparaciones transversales de diferentes cohortes de edad evaluadas al mismo tiempo (por ejemplo, las personas de 30 años comparadas con las de 50 que se evaluaron en 1977) y comparaciones de cohortes de edad específica evaluadas en momentos distintos (verbigracia, los treintañeros en 1963 y los treintañeros en 1984). Mediante el análisis estadístico de los datos de las distintas comparaciones fue posible diferenciar los cambios en la ejecución asociados con la edad de los relacionados, las modificaciones culturales de una sociedad particular y los vinculados a la práctica y experiencia individuales. Como se muestra en la figura 12.2, los resultados demostraron que, en casi todas las funciones, la declinación por la edad empieza más tarde y tiende a ser menos pronunciada de lo que planteaban las comparaciones transversales tradicionales.

Además de la aproximación amplia y sistemática al cuestionado efecto de la edad sobre la ejecución, el estudio investigó otras cuestiones relacionadas. Por ejemplo, se descubrió que los cambios con la edad varían según la función intelectual evaluada, como la habilidad verbal, la numérica y la velocidad perceptual (véase la figura 12.2), de ahí que una medida global de inteligencia (como el CI) puede haber confundido y oscurecido tanto los aumentos como las disminuciones en la habilidad. Se dedicó también mucha atención a las causas de los cambios, especialmente las disminuciones en la ejecución encontradas en los más viejos. Entre las principales variables asociadas con la disminución de la ejecución se hallan la mala salud, ciertas enfermedades, la inactividad, la falta de práctica de algunas funciones y diversas condiciones de personalidad, como la disminución de la motivación y una menor flexibilidad. Como resultado de estos descubrimientos, se elaboraron procedimientos de intervención en un esfuerzo por detener o revertir los decrementos en las capacidades, observados en el envejecimiento. El programa actual se dirige en particular al desarrollo de esos programas de remedio (véase K. W. Schaie, 1994; Schaie y Hertzog, 1986).

Diferencias individuales y edad. Además del descubrimiento básico de que los decrementos con la edad son menores y se presentan más tarde de lo que se había supuesto, la investigación actual revela considerables diferencias individuales en habili-

[3] Al principio, las pruebas incluían la Batería de Habilidades Mentales Primarias desarrollada por Thurstone mediante el análisis factorial de los aspectos de la inteligencia generalmente reconocidos (véase el capítulo 11). Después se preparó otra versión, el Test de Habilidades Mentales para los Adultos de Schaie-Thurstone, para emplearla en este proyecto (K. W. Schaie, 1988b). A la postre, las calificaciones ya no se reportaron para pruebas únicas, sino para constructos de modo que proporcionaran resultados más estables y generalizables.

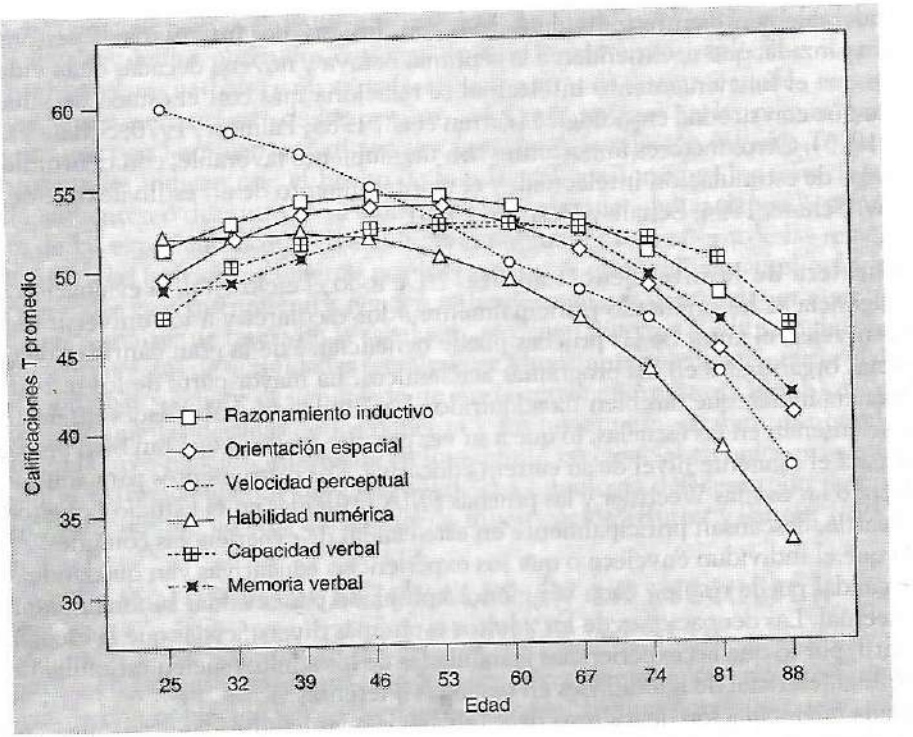


Figura 12.2. Estimaciones longitudinales de las calificaciones promedio del factor para los constructos de habilidad. (Tomado de K. W. Schaie, 1994, p. 308.) Nota: Datos acumulados de siete años de retests aplicados a las mismas personas.

(Copyright © 1994 por la American Psychological Association. Reproducido con autorización.)

dad en todos los niveles de edad. Cualquier generalización, sea que tenga que ver con el decremento con la edad o con diferencias de cohorte, debe ser matizada con el reconocimiento de la amplia variabilidad individual que se encuentra en todas las situaciones. Las diferencias individuales en cualquier nivel de edad son mucho mayores que las diferencias promedio entre los niveles de edad. En consecuencia, hay un considerable traslape en las distribuciones de las calificaciones obtenidas por individuos de edades distintas, lo que significa sencillamente que es posible encontrar un mayor número de ancianos cuya ejecución iguala la de los más jóvenes. Más aún, los de mejor ejecución en los grupos de mayor edad superan a los peores de los grupos de menor edad. Este traslape no se limita a los niveles adyacentes de edad, pues los rangos de ejecución siguen traslapados cuando se comparan grupos extremos, de modo que hay quienes a los 80 años superan a algunos veinteañeros.

Sin embargo, todavía más relevante para el tema es que los *cambios* que ocurren con el envejecimiento varían con el individuo. Esto supone, por ejemplo, que entre los 50 y los 60 años de edad algunos evidencian un declive, otros quizá no muestren cambios apreciables y otros más muestran un incremento en la ejecución en la prueba. La cantidad del cambio, sea una disminución o un aumento, también ha de variar

considerablemente entre individuos. Más aún, los estudios intensivos de personas de edad avanzada, que se extienden a la séptima, octava y novena décadas de la vida, indican que el funcionamiento intelectual se relaciona más con el estado de salud del sujeto que con su edad cronológica (Birren *et al.*, 1983; Palmore, 1970; Schaie y Gribbin, 1975). Otros factores importantes son un ambiente favorable, con oportunidades variadas de estimulación intelectual, y el mantenimiento de un estilo flexible de vida (K. W. Schaie, 1994; Schaie y Hertzog, 1986).

Naturaleza de la inteligencia adulta. De todo el ciclo vital, la evaluación de la inteligencia se ha orientado principalmente a los escolares y a los universitarios. A estos niveles, el autor de las pruebas puede beneficiarse de la gran cantidad de experiencias organizadas en los programas académicos. La mayor parte de los tests de inteligencia miden qué tan bien ha adquirido el individuo las habilidades intelectuales que se enseñan en las escuelas, lo que a su vez permite predecir qué tan bien preparado está para el siguiente nivel de su carrera educativa. Los instrumentos para adultos, incluyendo las escalas Wechsler y las pruebas PMA utilizadas en el Estudio Longitudinal de Seattle, descansan principalmente en este caudal de experiencias comunes. A medida que el individuo envejece o que sus experiencias educativas van quedando atrás, este caudal puede resultar cada vez menos apropiado para evaluar su funcionamiento intelectual. Las ocupaciones de los adultos están más diversificadas que la escolaridad infantil, por lo que las experiencias acumuladas de los adultos suelen estimular un desarrollo diferencial de habilidades en personas diferentes.

Dada la vinculación de los tests de inteligencia a las habilidades académicas, no resulta sorprendente que incluso los primeros estudios de adultos hayan mostrado mayores incrementos en las calificaciones con la edad entre quienes habían continuado por más tiempo su educación (D. P. Campbell, 1965; Härnqvist, 1968; Husén, 1951; Lorge, 1945; Owens, 1953). De igual modo, es probable que las personas cuyas ocupaciones tienen un contenido más "académico", que exigen la operación de habilidades verbales y numéricas, mantengan su nivel de ejecución o muestren mejoras con los años en las calificaciones que obtienen en los tests de inteligencia, mientras que aquellos cuyas ocupaciones requieren actividades mecánicas o relaciones personales pueden mostrar una pérdida. M. Williams (1960) comparó la ejecución de 100 personas, cuyas edades fluctuaban entre 65 y más de 90, en una serie de pruebas verbales y no verbales y obtuvo algunos datos interesantes que apoyan la hipótesis mencionada. Se encontraron correspondencias más bien sorprendentes entre la ocupación del individuo y su ejecución relativa en las dos clases de tareas. Las investigaciones longitudinales con adultos también han hallado relaciones interesantes entre los cambios en el CI total y ciertos reactivos del inventario biográfico (Charles y James, 1964; Owens, 1966).

Cada época y lugar fomentan el desarrollo de habilidades apropiadas a sus necesidades características que, a lo largo del ciclo vital, difieren para el infante, el escolar, el adulto en diferentes ocupaciones y el septuagenario retirado (Baltes, Reese y Lipsitt, 1980). Como demostración temprana de las implicaciones de este hecho para la evaluación de la inteligencia, Demming y Pressey (1957) empezaron con un análisis de tareas de las funciones habituales del adulto realizado mediante encuestas informales del material de lectura y del informe de los problemas y las actividades cotidianas. Sobre esta base, prepararon las formas preliminares de unas 20 pruebas "inherentes" a la

vejez. Las pruebas ponían de relieve la información práctica, el juicio y la percepción social. Los resultados obtenidos con tres de esos instrumentos, aplicados junto con pruebas estándares verbales y no verbales para muestras de distintas edades, mostraron que las personas de mayor edad superaban a los más jóvenes en las nuevas pruebas, mientras que con las pruebas tradicionales se sostenía la relación inversa. Todas estas investigaciones indican que el hecho de que las calificaciones en los tests de inteligencia aumenten o disminuyan al aumentar la edad en la adultez depende principalmente de las experiencias que viva el individuo durante esos años y de las relaciones entre éstas y las funciones cubiertas por las pruebas. A un nivel más amplio, los resultados de todos esos instrumentos pueden entenderse mejor dentro de un marco contextual. Los análisis de tareas de la conducta adulta en diversos medios, incluyendo el laboral y otras áreas de actividades socialmente importantes, debe mejorar la elaboración de pruebas y aclarar el significado de sus resultados (Anastasi, 1986b). Es digno de mención que las décadas de los ochenta y los noventa mostraran un notable aumento en la investigación interesada en los adultos, en especial en quienes se encuentran en el extremo superior del ciclo vital.[4] La psicología del desarrollo durante el ciclo vital representa un área floreciente de investigación (Rutter y Rutter, 1993).

CAMBIOS POBLACIONALES EN LA EJECUCIÓN EN LOS TESTS DE INTELIGENCIA

Aumento en las puntuaciones. ¿Qué sucede con la ejecución en los tests de inteligencia de una población luego de largos periodos? Es una pregunta que ya hemos encontrado en varias ocasiones. En la sección precedente vimos que la ejecución promedio en los tests de inteligencia se incrementó al aumentar el nivel educativo de la población adulta con el tiempo. Como resultado, los miembros más viejos de la muestra normativa, que en promedio tenían menos escolaridad, obtuvieron puntuaciones menores que los más jóvenes. Un fenómeno similar se halla en la muestra de estandarización de las pruebas para niños. En las reestandarizaciones sucesivas del Stanford-Binet y el WISC, la ejecución de las últimas muestras de estandarización fue considerablemente mejor que la de las muestras anteriores. En consecuencia, cualquier niño obtendría un CI menor si fuera evaluado con la edición revisada del que recibiría con la edición anterior, debido sencillamente a que su ejecución sería evaluada conforme normas más altas. El alto nivel educativo de los padres de los niños evaluados en la última muestra fue una de las condiciones mencionadas para explicar este aumento en la inteligencia evaluada.

Esta clase de comparación puede designarse como el estudio longitudinal de poblaciones. La aplicación común del método longitudinal en la investigación psicológica incluye la evaluación repetida de los mismos individuos en el tiempo; sin embargo, en el estudio longitudinal de poblaciones éstas son muestreadas en diferentes periodos. La comparación se hace entre cohortes de individuos nacidos en momentos distintos pero evaluados a las mismas edades.[5] Varias investigaciones a largo plazo conducidas

[4] Véase, por ejemplo, Bengtson y Schaie (1989), Birren y Bengtson (1988), Birren y Schaie (1991), Craik y Salthouse (1992), Fiske y Chiriboga (1990), Haplip y Panek (1993), Kausler (1994), Nadien (1989), Sonderegger (1992), Schulz y Ewen (1993) y Willis y Schaie (1986).

[5] Una aplicación especial de este procedimiento general puede observarse en la comparación de las series de tiempos incorporadas por K. W. Schaie (1965) en su diseño secuencial de cohortes.

durante la primera mitad del siglo XX revelaron un aumento en la inteligencia de la población de acuerdo con los tests estandarizados de inteligencia (Anastasi, 1985d, pp. 126–130). Al aumentar la alfabetización y los niveles educativos y presentarse otros cambios culturales, se hizo evidente que la inteligencia promedio de la población general en todas las edades mostraba un incremento estable durante varias décadas.

Estos estudios comparativos han empleado diversos procedimientos. Uno de ellos consiste en aplicar la misma prueba luego de cierto tiempo, como se hizo en los estudios de niños escoceses de 11 años de edad en 1932 y en 1947 (Scottish Council, 1949). Otro consiste en aplicar dos pruebas a una muestra representativa para establecer la correspondencia entre los dos conjuntos de puntuaciones y, por consecuencia, "traducir" la ejecución de una prueba a la otra. Esto fue lo que se hizo durante la Primera y la Segunda Guerras Mundiales al comparar la ejecución de los soldados del ejército estadounidense, que habían sido evaluados respectivamente con la Prueba Alfa del Ejército y con la Prueba de Clasificación General del Ejército (*Army General Classification Test*, Tuddenham, 1948). La tercera aproximación (técnicamente más adecuada) se basa en el establecimiento de una escala absoluta de calificación, libre de muestra, mediante el uso de reactivos ancla, como se hizo con las pruebas de la Junta Universitaria. La aplicación de la teoría de respuesta al ítem (TRI, capítulo 7) representa un perfeccionamiento posterior de esta aproximación.

Disminución de las puntuaciones. Que las puntuaciones que las poblaciones obtienen en los tests de inteligencia aumenten, disminuyan o permanezcan estables depende de muchas condiciones. Un factor importante es el *periodo* cubierto con los concomitantes cambios culturales. La *edad* de los examinados también supone una diferencia ya que, por ejemplo, un aumento en el nivel educativo de la población influirá directamente en la ejecución de los adultos en la prueba, pero en el caso de los niños lo hará sólo de manera indirecta porque las muestras comparadas han tenido la misma cantidad de educación en el momento de la evaluación. Por otro lado, también es importante considerar cualquier cambio en el *grado de selección* ocurrido en diferentes periodos, particularmente al examinar una subpoblación seleccionada. Por ejemplo, si en 1960 asistió a la secundaria una proporción mayor de la población que en 1910, como desde luego es el caso, entonces los estudiantes de secundaria de 1910 representan una muestra más seleccionada de la población general de su época que los que lo hicieron en 1960. Las aparentes inconsistencias entre el aumento y la disminución de las puntuaciones pueden deberse a la naturaleza de las pruebas empleadas, a la subpoblación evaluada (digamos, estudiantes de educación media que inician la universidad, la población general de adultos, niños de escuela primaria) o el periodo cubierto (por ejemplo, Flynn, 1984, 1987).

Un análisis de la reconocida disminución de las puntuaciones obtenidas con el Test de Aptitud Académica de la Junta Universitaria (*College Board's Scholastic Aptitude Test*) o SAT ilustra el número y la complejidad de las condiciones que pueden explicar el aumento o la disminución de la inteligencia evaluada en una población (Donlon, 1984, pp. 188–191; Wirtz, 1977). Entre 1963 y 1977, la puntuación verbal promedio del SAT cayó de 478 a 429, y la de matemáticas disminuyó de 502 a 470. En un esfuerzo por comprender esta disminución que permaneció estable durante 14

años, una comisión especial encargó a expertos en diversas áreas la realización de 38 estudios y consideró una cantidad impresionante de hipótesis causales.

Una conclusión importante a la que llegó la comisión fue que el patrón causal variaba de la primera a la segunda mitad del periodo de 14 años. Durante los primeros siete años, la disminución de las puntuaciones obedecía sobre todo a un cambio en la composición del grupo que presentaba el SAT. Debido a un incremento continuo observado en este periodo en la proporción de graduados de secundaria que asistían a la universidad, el muestreo se hizo progresivamente menos seleccionado en cuanto a las habilidades cognoscitivas medidas por la prueba; sin embargo, durante el segundo periodo de siete años la población universitaria se estabilizó y la selección del muestreo explicaba una parte mucho menor de la disminución en las puntuaciones. Para este periodo la explicación tendría que buscarse en las condiciones prevalecientes en el hogar, la escuela y el conjunto de la sociedad. La comisión observó que los datos disponibles no permitían determinar la contribución relativa de los distintos cambios culturales a la disminución de las puntuaciones. Empero, entre los muchos factores citados como probablemente significativos se encuentran un menor énfasis en los criterios académicos, la inflación en las puntuaciones y en las promociones automáticas, la disminución de los trabajos asignados para hacer en casa, un mayor ausentismo escolar, una menor atención al dominio de habilidades y conocimiento, el aumento excesivo en el tiempo dedicado a la televisión y las revueltas sociales del periodo que competían por la atención de los estudiantes.[6]

Un análisis subsecuente (Turnbull, 1985) planteó que la disminución en las puntuaciones del SAT durante el segundo periodo de siete años puede haber sido en parte una consecuencia demorada e indirecta del cambio en la composición de los primeros siete años. Como una mayor proporción de estudiantes mal preparados permanecían en secundaria (y solicitaban su admisión al bachillerato), muchos de los cambios en las condiciones escolares pueden considerarse respuestas adaptativas de las escuelas a la mayor diversidad de sus estudiantes. Estas respuestas, que tendían a disminuir el nivel de exigencia del programa escolar, se manifiestan en la inflación de las puntuaciones, la proliferación de cursos opcionales en las áreas vocacionales y generales, la simplificación de los textos y la reducción de las tareas. De ese modo, los cambios en las poblaciones estudiantiles condujeron a modificaciones en los programas educativos que a su vez dieron lugar a la disminución de las puntuaciones. Esta hipótesis es congruente con el descubrimiento de que hubo menos puntuaciones altas en los setenta. Las puntuaciones de los estudiantes que ocuparon los primeros lugares en sus grupos continuaron cayendo, mientras que las de los estudiantes que ocupaban los últimos lugares se estabilizaron o mejoraron. Más aún, a finales de los setenta y durante los ochenta se encontraron evidencias de un punto de cambio en el nivel de exigencia académica de las escuelas, lo que se refleja en un aumento en las puntuaciones del SAT.

Recapitulación. Un estudio de las ganancias informadas en 14 naciones en la ejecución en tests de inteligencia resalta los problemas metodológicos enfrentados al tratar de evaluar los cambios poblacionales (Flynn, 1987). Los descubrimientos revelaron

[6] Aunque investigados más exhaustivamente por el SAT, en otros exámenes de admisión universitaria se observaron disminuciones similares en las calificaciones, como en las del programa de Evaluación de Universidades Estadounidenses (*American College Testing*, ACT) y en los niveles de escuelas primaria y secundaria.

tantas inconsistencias e incompatibilidades, y sus explicaciones eran tan escurridizas, que llevaron al autor a la conclusión de que los tests de inteligencia en realidad no medían la inteligencia. Los esfuerzos por medir los cambios poblacionales aún siguen bajo investigación y aún queda por desarrollar un procedimiento óptimo. El lector puede encontrar un excelente tratamiento de las cuestiones relevantes de procedimiento en la publicación de una conferencia dedicada al tema, editada por Collins y Horn (1991).

Para comprender los descubrimientos de aumentos y disminuciones en la ejecución de las poblaciones se requieren varios tipos de información relacionada. Primero, es preciso identificar las pruebas particulares que se aplicaron, con especial atención a su cobertura de procesos y contenidos específicos (por ejemplo, velocidad perceptual, memoria, comprensión verbal y orientación espacial), así como la fuente de sus normas. Segundo, es preciso informar de las fechas del test y el retest. Tercero, deben incluirse los datos pertinentes de las poblaciones muestreadas y cualquier cambio de muestreo en el retest, como una pérdida de los individuos que inicialmente tenían una ejecución muy buena o muy mala en las muestras sucesivas. Cuarto, también debe tenerse acceso a la información sobre cualquier cambio cultural importante que pueda afectar la población estudiada, lo que puede referirse a la cantidad y naturaleza de la educación, a desarrollos en las comunicaciones o transportes que puedan influir en los contactos entre culturas o cualquier otra experiencia que pueda tener un efecto (en cantidad o dirección) en el progreso intelectual de los individuos.

Las posibles confusiones que pueden interrumpir el estudio de los cambios en diferentes poblaciones también incluyen el inadecuado uso popular del término "inteligencia" (y especialmente del "CI") para representar una propiedad unitaria e identificable del organismo (véase, por ejemplo, Flynn, 1987). Si se considera a la inteligencia como la combinación de habilidades requerida para el funcionamiento eficaz y el progreso en un contexto particular (véase el capítulo 11), entonces es necesario contar con información sobre las principales condiciones citadas antes para hacer una interpretación apropiada de la ejecución del individuo en un test de inteligencia o de la ejecución promedio de poblaciones en distintos momentos y lugares. Cada vez se reconoce más la gran cantidad de problemas técnicos que se enfrentan al tratar de medir el cambio y la diversidad de procedimientos para evaluarlo en contextos y con propósitos diferentes (véase especialmente Gottman, 1995).

DIVERSIDAD CULTURAL

En nuestro libro, examinamos desde distintos ángulos la aplicación de pruebas a personas con diversos antecedentes culturales. El capítulo 18 se interesa en las consideraciones éticas y sociales de dicha evaluación, en especial en lo que atañe a los grupos minoritarios en las culturas nacionales En los capítulos 6 y 7 estudiamos los problemas técnicos relacionados con el sesgo del instrumento y de los reactivos; y en el capítulo 9 revisamos algunas de las pruebas originalmente diseñadas para diversas aplicaciones transculturales. En esta sección, presentaremos algunos de los principales temas teóricos sobre la función de la cultura en la conducta, con referencia especial al uso e interpretación de las calificaciones de los tests de inteligencia.

El campo de la psicología de la cultura. Las tres últimas décadas del siglo XX han presenciado un notable crecimiento en la investigación y los trabajos publicados sobre la psicología de la cultura (Berman, 1990; Irvine y Berry, 1988). Se han dedicado incluso conferencias internacionales al tema (véase, por ejemplo, Brislin, 1993; Cronbach y Drenth, 1972; Manoleas, 1995). El campo atañe principalmente a las diferencias conductuales entre grupos que fueron educados y que funcionan en diversos contextos culturales identificables, contextos que pueden ser tan estrechos como un vecindario o una ciudad o tan amplios como una nación o un continente. Más aún, algunas culturas bien identificadas, como los hispanos en los Estados Unidos, presentan subculturas, como los mexicanos, los cubanos, los puertorriqueños y los centro y sudamericanos, que son lo bastante diferentes como para requerir una identificación particular al tratar de comprender la conducta de los individuos (véase, por ejemplo, Geisinger, 1992; Marín y Marín, 1991).

La función de la cultura en la conducta humana puede concebirse como una forma de especificidad de dominio, similar a la que reconoce la psicología cognoscitiva, como vimos en el capítulo 11. Empezando con el análisis de los procesos psicológicos básicos, como el aprendizaje, el recuerdo, la resolución de problemas y la emoción, los psicólogos cognoscitivistas pronto descubrieron que estos procesos se manifestaban como conducta específica a un dominio; por ejemplo, la memoria (o la solución de problemas o el razonamiento) pueden variar considerablemente dependiendo de si se juega ajedrez, se resuelven problemas matemáticos o se escribe un ensayo.

La psicología de la cultura comenzó con el estudio del comportamiento en culturas muy diversas y hasta entonces desconocidas que sin duda eran distintas a la cultura del investigador. Rápidamente se convirtió en una exploración sistemática de las diferentes experiencias de personas criadas en culturas disímiles. En esencia, el campo actual de esta área representa el reconocimiento de la especificidad cultural de la conducta humana, de que los procesos psicológicos básicos pueden dar por resultado ejecuciones, actitudes, autoconceptos y visiones del mundo muy diferentes en los miembros de diversas culturas (L. L. Adler y Gielen, 1994; Berry, Poortinga, Segall y Dasen, 1992; Díaz-Guerrero, 1990; Shweder y Sullivan, 1993). La contribución de la cultura se reconoce e integra cada vez más en todos los campos de la psicología, desde la investigación y la teoría sobre el desarrollo en el ciclo vital, la conducta social, la emoción o el pensamiento[7] por un lado, hasta la práctica de la psicología industrial, organizacional, clínica o de la consejería, por el otro.[8]

La mayor importancia que se concede a la función de los factores culturales en todos los campos de la psicología quedó demostrada en el programa de la convención anual realizada en 1994 por la Asociación Estadounidense de Psicología (*American Psychological Association*, APA, 1994), además de un taller dedicado a la "sensibilidad cultural en la evaluación y la intervención". Las conferencias magistrales se dedicaron al tema general de las "perspectivas internacionales de la psicología transcultural". En esta serie, seis psicólogos con reconocimiento internacional debatieron los aspectos

[7] Véase, por ejemplo, Gormly y Brodzinsky (1993), Kitayama y Marcus (1994), Mistry y Rogoff (1985), Nugent, Lester y Brazelton (1991), Rogoff (1990), Rogoff y Chavajay (1995), Smith y Bond (1993) y Topping, Crowell y Kobayashi (1989).

[8] Véase, por ejemplo, Freilich, Raybeck y Savishinsky (1991), Pedersen (1987), Pedersen e Ivey (1993) y Triandis, Dunnette y Hough (1994).

del tema. Las conferencias fueron parte de la Miniconvención Temática sobre la "Diversidad cultural: El futuro de los Estados Unidos", que se distribuyó durante los cinco días que duró la convención de la APA. Por si no bastara, la aparición de una nueva publicación, *Culture and Psychology* (1995), es otro indicador del rápido crecimiento de la orientación transcultural de la psicología.

Diferencias y desventajas culturales. Cuando los psicólogos empezaron a preparar instrumentos para la evaluación transcultural en la primera parte del siglo XX, esperaban que al menos en teoría sería posible medir el "potencial intelectual hereditario" independientemente del impacto que pudieran tener las experiencias culturales. Se pensaba que la conducta del individuo estaba cubierta con una especie de barniz cultural cuya penetración se convirtió en el objetivo de lo que empezó a llamarse pruebas "libres de cultura". El desarrollo de la genética y la psicología ha demostrado la falacia de dicho concepto. Ahora reconocemos que los factores hereditarios y ambientales operan a la vez en todas las etapas del desarrollo del organismo y que sus efectos se encuentran inextricablemente vinculados en la conducta resultante. En el ser humano, la cultura permea de hecho todos los contactos ambientales, y dado que toda la conducta está influida por el medio cultural en el que el individuo se desarrolla y que las pruebas psicológicas son apenas muestras de conducta, el efecto de la cultura debe verse reflejado por fuerza en la ejecución del instrumento. Es entonces inútil tratar de construir una prueba *libre* de las influencias culturales. Más tarde, el objetivo cambió a la preparación de pruebas que sólo consideraran las experiencias que fueran *comunes* a las diferentes culturas. Por esta razón el término "libre de cultura" fue sustituido por términos como "cultura común", "culturalmente justas" y "transculturales".

No obstante, no hay una sola prueba que pueda aplicarse de modo universal o que sea "justa" para todos, por lo que existen tantas variantes de pruebas culturalmente justas como parámetros en que las culturas difieren. Un instrumento sin lectura puede ser culturalmente justo en una situación, en otra puede serlo una prueba sin lenguaje, en una más lo sería una prueba de ejecución, y en la cuarta la traducción adaptada de una prueba verbal. Aunque las variantes de las pruebas transculturales disponibles no son intercambiables, sí son útiles para distintas comparaciones transculturales. Más aún, es poco probable que cualquier prueba pueda ser "justa" para más de un grupo cultural, en especial cuando las culturas son muy disímiles. Aunque las reducen, las pruebas transculturales no pueden eliminar por completo las diferencias culturales en la ejecución de los instrumentos, pues cada uno tiende a favorecer a las personas de la cultura en el que se elaboró. El simple uso de lápiz y papel o la presentación de tareas abstractas que no tienen significado práctico inmediato resultarán ventajosos para ciertos grupos, pero para otros representarán una desventaja. Los factores emocionales y motivacionales también influyen en los tests de ejecución. Entre las condiciones relevantes que difieren de una cultura a otra pueden mencionarse el interés intrínseco del contenido del instrumento, el *rapport* establecido con el examinado, la motivación para hacer un buen trabajo en la prueba, el deseo de superar a los demás y el estilo, individual o cooperativo, de resolver los problemas.

Las diferencias culturales se convierten en desventajas cuando el individuo se traslada de la cultura o subcultura en la que se educó y se esfuerza por funcionar, competir o tener éxito en otra. No obstante, desde un punto de vista más amplio son justamente

esos contactos e intercambios entre culturas los que estimulan el avance de las civilizaciones. El aislamiento cultural, aunque tal vez sea más cómodo para los individuos, lleva al estancamiento social.

Un concepto relacionado es el de privación cultural. Aunque el término se ha empleado en muchos sentidos diferentes, Feuerstein (1980, 1991; Feuerstein y Feuerstein, 1991) le dio un significado especial y lo convirtió en el punto central de su programa de capacitación cognoscitiva. Consideraba a la privación cultural como un estado de reducción de la modificación cognoscitiva producida por la carencia de una *experiencia de aprendizaje mediado*. La transmisión del conocimiento acumulado por la cultura de una generación a la siguiente es característica de la especie humana. En este proceso, los padres u otra persona encargada del cuidado del niño actúan como *agentes de mediación* al seleccionar y organizar los estímulos que le presentan al niño. Feuerstein considera que el aprendizaje mediado es esencial para el desarrollo cognoscitivo en la medida en que fomenta la disposición para aprender, orientaciones y otros patrones de conducta que facilitan el aprendizaje subsecuente. Los niños que, por la razón que sea, no logran experimentar este aprendizaje mediado carecen de los requisitos para el funcionamiento cognoscitivo para el nivel superior. En contraste, quienes han tenido tales experiencias en su propia cultura han adquirido las habilidades y hábitos que se precisan para la modificación y pueden adaptarse a las exigencias de una nueva cultura luego de un breve periodo de transición. También es probable que en las naciones en desarrollo los psicólogos que forman parte de la cultura nacional lleguen a elaborar y utilizar instrumentos apropiados para su cultura.

Desde un punto de vista diferente, la simple existencia de *estereotipos culturales* puede influir directamente en la ejecución del individuo en las pruebas (Steele, Spencer y Aronson, 1995). En un programa de investigación a largo plazo se encontró evidencias de que el conocimiento de los estereotipos existentes tiene un efecto en la motivación y las actitudes de algunos examinados hacia la prueba por las distracciones, el autoconcepto, el menor esfuerzo y las pocas expectativas de un desempeño exitoso. Esta reacción se conoce como *vulnerabilidad estereotipada*, y se ha descubierto que influye en las calificaciones de las pruebas en comparaciones tanto de género como étnicas. Para ser eficaces, los programas de acción afirmativa necesitan ofrecer algo más que entrenamiento remedial. Al menos en algunos casos, las expectativas de ejecución en áreas específicas, como la verbal, la matemática o la espacial, pueden requerir de atención especial para prevenir el fracaso de quienes de otra manera habrían mostrado un desempeño normal o de nivel superior.

El lenguaje en la evaluación transcultural. Las pruebas transculturales más tradicionales utilizaban un contenido no verbal con la esperanza de obtener una medida más imparcial de las mismas funciones intelectuales medidas con instrumentos de inteligencia verbal (véase el capítulo 9). Las dos suposiciones que fundamentan este planteamiento son cuestionables. Primero, no puede suponerse que los instrumentos no verbales miden las mismas funciones que las pruebas verbales, por similares que puedan parecer. Una prueba de analogías espaciales no es sencillamente una versión no verbal de un instrumento de analogías verbales. Algunos de los primeros tests sin lenguaje, como el Beta del ejército estadounidense, tenían una fuerte carga de habilidades de visualización espacial y perceptuales muy poco relacionadas con las verbales

y numéricas. Incluso en pruebas como las Matrices Progresivas de Raven y otros instrumentos no verbales diseñados deliberadamente para tocar el razonamiento y la conceptuación abstracta, los análisis factoriales han revelado una contribución importante de los factores no verbales a la varianza de las puntuaciones (por ejemplo, R. S. Das, 1963). La investigación más reciente de la psicología cognoscitiva, que ha demostrado repetidamente la *especificidad de contenido* de los procesos de pensamiento, corrobora tales descubrimientos. Las estrategias y habilidades requeridas para la resolución de problemas se adquieren mientras se responde a dominios particulares de contenido y en contextos particulares (véase el capítulo 11).

Desde otro ángulo, una gran cantidad de evidencias plantea que las pruebas sin lenguaje pueden tener una mayor carga cultural que las pruebas con lenguaje. Las investigaciones realizadas en diversos países con una amplia variedad de grupos culturales encontraron más diferencias grupales en las pruebas de desempeño y otros instrumentos no verbales que en las pruebas verbales (Irvine, 1969a, 1969b; Jensen, 1968; Ortar, 1963, 1972; Trimble, Lonner y Boucher, 1983; Vernon, 1969). Se cuenta también con evidencias de que las pruebas con figuras pueden ser más susceptibles a los efectos de la capacitación que las pruebas verbales y numéricas (Irvine, 1983). El propio uso de una representación pictórica puede ser inadecuado en culturas en las que no se acostumbra el dibujo representativo. Una reproducción bidimensional de un objeto no es una réplica exacta del original; sencillamente presenta algunas señales que, como resultado de la experiencia previa, llevan a la percepción del objeto. Si las señales se reducen de manera considerable, como en la imagen de la cara para representar a toda la persona, y si se carece de la experiencia previa necesaria, puede ocurrir que no tenga lugar la percepción correcta. Se cuenta ahora con considerables datos empíricos que revelan la existencia de marcadas diferencias en la percepción de imágenes en personas de diferentes culturas (R. J. Miller, 1973; Segall, Campbell y Herskovits, 1966).

Todavía desde otro punto de vista, a menudo las pruebas no verbales, espaciales perceptuales, requieren de procesos de pensamiento relativamente abstractos y estilos cognoscitivos analíticos que son característicos de la clase media de las culturas occidentales (J. W. Berry, 1972; R. A. Cohen, 1969). Quienes crecieron en otros contextos culturales pueden estar menos acostumbrados a estos métodos de resolución de problemas. Las culturas difieren en el valor que asignan a la generalización y a la búsqueda de rasgos comunes en situaciones desiguales. En algunas, la conducta generalmente se vincula a contextos y situaciones y las respuestas a las preguntas pueden depender de quién la plantea y de su contenido (Cole y Bruner, 1971; Goodnow, 1976; Neisser, 1976, 1979).

Por supuesto, no hay dificultades de procedimiento en la aplicación de una prueba verbal a culturas que hablan el mismo idioma, pero cuando las diferencias idiomáticas requieren la traducción de la prueba, pueden surgir problemas relacionados con la comparabilidad de las normas y la equivalencia de las calificaciones. Debe advertirse que rara vez basta con la simple traducción, por lo que en general se requiere de adaptación y revisión. El contenido de algunos reactivos puede ser más familiar para una cultura que para otra, y algunas diferencias todavía más sutiles pueden afectar significativamente la ejecución en la prueba; por ejemplo, la longitud relativa de las palabras o el sonido similar de diferentes vocablos de uno de los idiomas puede alterar la dificultad de lectura en ambas lenguas (Valencia y Rankin, 1985). Tampoco puede su-

ponerse que la versión original de una prueba y su traducción son equiparables porque las dos versiones pueden diferir de muchas maneras (Durán, 1989; Marín y Marín, 1991). Los *Estándares de Examinación* (AERA, APA, NCME, 1985, capítulo 13) especifican con claridad que la confiabilidad, la validez y las normas de la versión traducida deben establecerse independientemente para cualquier población en la que vaya a utilizarse el instrumento.

No olvidemos que es probable que los factores culturales que inciden en las respuestas a la prueba también influyan en el dominio más amplio de comportamiento que la prueba pretende muestrear; por ejemplo, en las culturas de habla inglesa el dominio inadecuado del idioma puede suponer una desventaja para el niño no sólo en un test de inteligencia, sino también en el trabajo escolar, los contactos con los compañeros y las actividades de juego, que a su vez pueden interferir con el desarrollo emocional e intelectual subsecuente. En el adulto, pueden restringir seriamente el desempeño laboral, las relaciones personales y otras actividades importantes. Podríamos citar muchos otros ejemplos similares de diferencias culturales. Algunas son diferenciales cognoscitivas, como problemas de lectura o estrategias inadecuadas para solucionar problemas abstractos; otras, actitudinales o motivacionales, como la falta de interés en las actividades intelectuales, la hostilidad a las figuras de autoridad, una baja motivación de logro o un autoconcepto pobre. Todas estas condiciones pueden mejorarse de maneras diversas, que van de la enseñanza del idioma y la alfabetización funcional a la consejería personal y la psicoterapia. Es probable que todos éstos tengan un efecto tanto en la ejecución en las pruebas como en las actividades cotidianas de niños y adultos.

La importancia del lenguaje en la ejecución en las prueba y en las actividades cotidianas ha llevado al desarrollo de pruebas de destreza lingüística en el idioma original y el adoptado. En los Estados Unidos, la mayor parte de las pruebas disponibles se manejan en inglés y en español. Durán (1989, pp. 574–577) ofrece una lista de los ejemplos de pruebas, así como las referencias a la investigación, el desarrollo y la evaluación de las mismas. En De Groot y Barry (1993) puede encontrar un tratamiento más amplio del bilingüismo.

La situación de prueba. La rápida difusión de los contactos transculturales que se observa en el mundo actual incrementa la probabilidad de que las pruebas se apliquen a sujetos de culturas diferentes. Cada examinador puede anticipar la evaluación de una o más personas de una cultura distinta a la suya, de ahí que su capacitación deba incluir cierto conocimiento de una o más culturas distintas y prestar especial atención a los probables efectos de la cultura sobre el desarrollo conductual de los individuos. Todavía más importantes son los probables efectos de las diferencias en la respuesta del examinado a la evaluación. Algunos ejemplos de las fuentes de conducta diferencial ante la evaluación incluyen las variaciones en el autoconcepto de la persona, su visión del mundo, su grado de autorrevelación y los hábitos de resolver los problemas por separado o en grupos.

Como explicamos en la primera parte del capítulo, eliminar de un instrumento las partes que se suponen especialmente difíciles por los antecedentes culturales del individuo sólo reducirá su validez predictiva e impedirá informar al examinado sobre las áreas que debe fortalecer para tener un desempeño eficaz en el contexto anticipado.

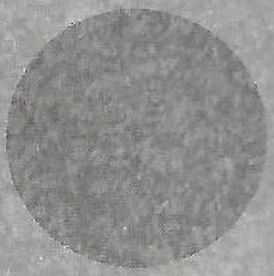
En consecuencia, la aproximación tradicional está siendo reemplazada por una solución que cambia el énfasis a la ~~*conducta del evaluador*~~ en la situación de prueba.

Las décadas de los ochenta y los noventa han producido una serie de manuales, guías, revistas, artículos y otras publicaciones relacionadas con la capacitación y la conducta apropiada de los evaluadores al examinar a los miembros de otras culturas (Atkinson, Morten y Sue, 1993; Myers, Wohlford, Guzmán y Echemendia, 1991; Stricker *et al.*, 1990). Algunas abordan en particular la evaluación de estudiantes (Samuda, Kong, Cummins, Lewis y Pascual-Leone, 1991); otras se concentran en el examen de los niños (Miller-Jones, 1989; Rogoff y Morelli, 1989), y otras más en alguna población cultural particular o en alguna de sus subdivisiones (Dana, 1984; Díaz Guerrero y Szalay, 1991), pero la inmensa mayoría tiene que ver con el problema general de la evaluación transcultural. Los trabajos fluctúan entre las directrices muy generales y concisas publicadas en 1991 por la Asociación Estadounidense de Psicología en un breve cuadernillo y que se reimprimieron en el número de 1993 del *American Psychologist* ("Guidelines", 1993) al exhaustivo y detallado tratamiento del libro de Dana (1993). Estas dos fuentes generales deben ser estudiadas con sumo cuidado por cualquiera que planee la aplicación de instrumentos. También es importante advertir que en Groot y Barry (1993) puede encontrarse una amplia revisión de las variedades, los problemas y las ventajas del bilingüismo.

En principio, la función especial del examinador en las evaluaciones transculturales incluye, primero, obtener en una sesión inicial con el examinado la información completa sobre su identidad cultural, el grado y el tipo de aculturación y las características de la cultura de origen que probablemente influyan en su desempeño en la prueba. Segundo, adaptar su conducta a las necesidades del examinado. A este respecto, el examinador debe considerar cómo debe presentar la prueba, cómo explicar su propósito y cómo motivar al examinado para que se desempeñe apropiadamente. Más aún, los propios modales del examinador deberían establecer una relación personal favorable con el examinado. Desde luego, el uso y la interpretación de las puntuaciones, así como la naturaleza de la realimentación y de quien la recibe deben tener en consideración los factores culturales.

Richard Dana (1993) ha elaborado un conjunto de cuestionarios para ayudar a los examinadores a obtener la información que requieren de los examinados. Algunos son de uso general, otros se desarrollaron para ciertas culturas (por ejemplo, los estadounidenses nativos o los de origen asiático) o para algunas de sus subculturas particulares. Como último comentario, agreguemos que el énfasis en la obtención de al menos *parte* de la información relevante sobre las experiencias del examinado para la comprensión y uso adecuados de las puntuaciones en las pruebas dirige la atención a una práctica que es deseable en la evaluación de *cualquier* persona, independientemente de las diferencias culturales.

C U A R T A
P A R T E



Evaluación de la personalidad

Inventarios autodescriptivos de personalidad

Aunque en ocasiones se emplea el término "personalidad" en un sentido más amplio, en la terminología psicométrica los "tests de personalidad" son instrumentos que se utilizan para medir las características emocionales, motivacionales, interpersonales y de actitudes, rasgos todos estos que se consideran distintos de las habilidades.[1] En los próximos cuatro capítulos, examinaremos los más importantes tests de personalidad clasificados según el método empleado para obtener los datos de los individuos. En este capítulo revisaremos los inventarios autodescriptivos de personalidad y en el siguiente las técnicas usadas para medir intereses y actitudes. Los instrumentos descritos en estos dos capítulos son sobre todo cuestionarios autodescriptivos, instrumentos de lápiz y papel que si bien son particularmente adecuados para la aplicación colectiva, muchos pueden utilizarse también en la evaluación individual. En el capítulo 15 analizaremos el uso de las técnicas proyectivas para evaluar las características de personalidad y en el 16 revisaremos diversas aproximaciones al tema, algunas de las cuales se encuentran aún en etapa experimental.

El número de tests de personalidad alcanza varios cientos, de los que son especialmente numerosos los inventarios de personalidad y las técnicas proyectivas. En este libro nos interesamos ante todo en los acercamientos que han sido explorados en la evaluación de la personalidad; por ello, y aunque se ha escrito varios libros dedicados en exclusiva al tema, en un intento ilustrativo haremos una breve descripción de algunas de las pruebas más conocidas de cada categoría. Para una revisión más detallada del

[1] El lector puede encontrar en L. R. Aiken (1993), Burger (1993) y Maddi (1989) más antecedentes sobre las aproximaciones actuales a la teoría e investigación de la personalidad.

tema, el lector puede recurrir a libros especializados (por ejemplo, Angleitner y Wiggins, 1986; Butcher, 1995; Lanyon y Goodstein, 1997; y J. S. Wiggins, 1973/1988).

En el desarrollo de los inventarios de personalidad se han seguido diversas aproximaciones para formular, reunir, seleccionar y agrupar los reactivos. Entre los procedimientos de mayor uso se encuentran los que se basan en la relevancia del contenido, la clave del criterio empírico, el análisis factorial y la teoría sobre la personalidad. En las siguientes secciones analizaremos e ilustraremos cada uno de esos planteamientos, aunque es importante anticipar que no se trata de técnicas alternativas o mutuamente excluyentes, pues todas pueden combinarse, teóricamente, en el desarrollo de un solo inventario de personalidad y, de hecho, en la mayor parte de los instrumentos actuales se utilizan dos o más de estos procedimientos.

Aunque algunos tests de personalidad se emplean como instrumentos colectivos de proyección, casi todos encuentran su principal aplicación en la clínica y la consejería, por esta razón los cuatro capítulos siguientes deberán considerarse en conjunción con los puntos de vista de los contextos especializados que veremos en el capítulo 17. En su estado actual de desarrollo, hay que considerar la mayor parte de los tests de personalidad como auxiliares en la evaluación individual o como instrumentos de investigación.

PROCEDIMIENTOS RELACIONADOS CON EL CONTENIDO

El prototipo de los inventarios autodescriptivos de personalidad fue la Hoja de Datos Personales de Woodworth (*Woodworth Personal Data Sheet*), instrumento desarrollado para usarse durante la Primera Guerra Mundial (véase la referencia en el capítulo 2) que en esencia constituyó una tentativa por estandarizar una entrevista psiquiátrica y adaptar el procedimiento para la evaluación masiva. Para ello, Woodworth recopiló de la bibliografía psiquiátrica y de conversaciones con especialistas de esta disciplina la información concerniente a los síntomas comunes de neuróticos y preneuróticos. Las preguntas del inventario se formularon en función de esos síntomas y trataban las desviaciones de la conducta como miedos anormales o fobias, obsesiones y compulsiones, pesadillas y otros trastornos del sueño, fatiga excesiva y otros síntomas psicosomáticos, sentimientos de irrealidad y perturbaciones motoras como tics y temblores. Si bien es cierto que en la elección final de los reactivos Woodworth utilizó ciertas herramientas estadísticas (que veremos en la próxima sección), también es evidente que en la elaboración y uso de este inventario puso el acento en la relevancia del contenido de sus reactivos, como lo indican las fuentes de las que se obtuvieron así como el reconocimiento común de ciertas clases de comportamiento como maladaptados. El legado de Woodworth a la elaboración de los inventarios modernos se aprecia en el hecho de que al preparar un nuevo banco de reactivos las instrucciones de quienes los escriben se basan en el análisis de contenido del área conductual que pretende evaluarse.

Un ejemplo actual de la aproximación relacionada con el contenido al desarrollo del inventario autodescriptivo es la Lista de Verificación de 90 Síntomas, Revisada (*Symptom Checklist-90-Revised, SCL-90-R*, Derogatis, 1994), un instrumento diseñado para identificar problemas psicológicos y síntomas de psicopatología, que consta de la breve descripción de 90 síntomas (por ejemplo, mal apetito, debilidad o vértigos). En una escala de cinco puntos se pide a los individuos que indiquen el grado al

que se han visto afligidos por cada uno de los problemas en los últimos siete días. La SCL-90-R proporciona normas separadas para adultos y adolescentes de ambos sexos no pacientes así como para pacientes psiquiátricos internados y de consulta externa. Sin embargo, algunas de las normas no son lo suficientemente representativas; por ejemplo, el grupo normativo de pacientes psiquiátricos internos estuvo conformado predominantemente por individuos de posición socioeconómica baja, y el grupo de adolescentes, sobre todo por personas de clase media y casi todas blancas.

Los reactivos de la SCL-90-R no sólo se asemejan a los de los instrumentos anteriores del tipo de lista de verificación en el hecho de que se seleccionaron sobre la base de la relevancia del contenido y la utilidad clínica, sino también en que algunos de ellos pueden rastrearse hasta la Hoja de Datos Personales de Woodworth mediante las escalas intermedias como la Lista de Verificación de Síntomas de Hopkins (*Hopkins Symptom Checklist*) y el Índice Médico de Cornell (*Cornell Medical Index*, Derogatis y Lazarus, 1994). Los reactivos se organizan en nueve dimensiones de psicopatología: somatización, depresión, ansiedad, hostilidad, psicoticismo, sensibilidad interpersonal, ansiedad fóbica, ideación paranoide y síntomas obsesivo-compulsivos. Los análisis factoriales de las escalas indican que están correlacionadas y que, por consecuencia, no son de mucha utilidad para el diagnóstico diferencial; no obstante, los índices globales derivados de la lista de verificación han demostrado ser indicadores confiables de la presencia y la gravedad de la psicopatología (Payne, 1985). La SCL-90-R y otros instrumentos relacionados, como el Inventario Breve de Síntomas (*Brief Symptom Inventory*), se utilizan mejor como parte de una batería, en la evaluación del cambio producido en el curso de una terapia y en la investigación sobre el efecto de diversos tratamientos.

La principal ventaja de la aproximación relacionada con el contenido para el desarrollo de los inventarios de personalidad estriba en la simplicidad y lo directo del método. No obstante que esas características hacen posible contar con instrumentos relativamente breves y económicos, su transparencia permite que los examinados hagan esfuerzos conscientes por manipular los resultados en mayor medida que otros métodos; además, en general estos instrumentos no poseen las características destinadas a prevenir o detectar los sesgos de respuesta que revisaremos más adelante (Bornstein, Rossner, Hill y Stepanian, 1994), por lo que no resulta aconsejable confiar exclusivamente en sus resultados como la base para la toma de decisiones.

CLAVE EMPÍRICA DE CRITERIO

Aproximación básica. El uso de *una clave empírica de criterio* se refiere al desarrollo de una clave de puntuación en términos de algún criterio externo. El procedimiento requiere la selección de los reactivos que han de conservarse y la asignación de los pesos de la puntuación para cada respuesta. En la elaboración de la Hoja de Datos Personales de Woodworth, algunas de las comprobaciones estadísticas aplicadas en la selección final de los reactivos señalaron el camino para la clave de criterio. Así, en este inventario no se conservó ningún reactivo si el 25 por ciento o más de los miembros de la muestra normal lo respondía en la dirección desfavorable por considerar que una

característica conductual que ocurre con tal frecuencia en una muestra esencialmente normal no puede ser indicador de anormalidad. En la selección de reactivos se utilizó también el método de grupos contrastados, por lo que sólo se conservaron los síntomas que en un grupo diagnosticado como psiconeurótico se informaban al menos dos veces más que en un grupo normal.

No obstante, a pesar del uso de dichas comprobaciones empíricas, las aproximaciones relacionadas con el contenido confían esencialmente en una interpretación literal o verídica de los reactivos del cuestionario. Se considera que la respuesta a cada pregunta es un indicador de la actual existencia o inexistencia del problema, creencia o conducta descrita por la pregunta. Por otro lado, en la clave empírica de criterio las respuestas se consideran diagnósticas o sintomáticas del criterio conductual con el que se encontraron asociadas.

En una descripción inicial de esa aproximación, Meehl (1945) escribía:

> . . . la forma más fructífera de considerar al inventario de tipo verbal de personalidad *no* es como una "autoclasificación" o autodescripción cuyo valor requiere suponer que la precisión en las observaciones que el examinado hace de sí mismo son exactas. Más bien, las respuestas a los reactivos de la prueba se consideran como un segmento intrínsecamente interesante de la conducta verbal, cuyo conocimiento puede ser más importante que cualquier conocimiento del material "fáctico" sobre el que el reactivo pretende investigar de manera superficial. De modo que, si un hipocondriaco dice que sufre "muchas jaquecas", el hecho que resulta de interés es que él lo *dice*. [p. 9]

Un inventario autodescriptivo consta de una serie de estímulos verbales estandarizados, cuyas respuestas se califican —en los casos en los que utilizan procedimientos de clave de criterio— en función de sus correlatos conductuales establecidos empíricamente, lo que implica que se les trata como a cualquier otra respuesta a una prueba psicológica. El hecho de que las respuestas correspondan a la *percepción* que el sujeto tiene de la realidad no altera dicha situación, sencillamente brinda una hipótesis para explicar la validez establecida de manera empírica de algunos reactivos.

Inventario Multifásico de Personalidad de Minnesota

El Inventario Multifásico de Personalidad de Minnesota (MMPI) constituye el ejemplo más destacado de la clave de criterio empírico en la elaboración de tests de personalidad. En años recientes, el MMPI ha sido revisado y reformulado en dos versiones, el MMPI-2 (Butcher, Dahlstrom, Graham, Tellegen y Kaemmer, 1989) y el MMPI para Adolescentes (MMPI-A, Butcher *et al.*, 1992). De cualquier modo, no es posible analizarlas sin referirse al MMPI original y a la función que cumplió en la historia de la evaluación de la personalidad. Aunque una descripción detallada del instrumento rebasa el alcance de este texto, debe advertirse que, durante casi medio siglo, el MMPI fue el test de personalidad de mayor uso y el más investigado.[2]

[2] Para una descripción más concisa del MMPI original, véanse las ediciones anteriores de este texto (por ejemplo, Anastasi, 1988b); una revisión más extensa puede encontrarse en los volúmenes clásicos de Dahlstrom, Welsh y Dahlstrom (1972, 1975).

Como instrumento, el MMPI fue, en muchos sentidos, víctima de su propio éxito. Concebido en la década de los treinta por Starke R. Hathaway, psicólogo clínico, y por J. Charles McKinley, neuropsiquiatra, durante los años cuarenta se publicó en una serie de artículos como auxiliar en el proceso del diagnóstico psiquiátrico.[3] Posteriormente, su efectividad para detectar psicopatologías y diferenciar entre las categorías nosológicas gruesas condujo a su utilización con distintos propósitos que superaban su objetivo original. Para la década de los sesenta, el MMPI estaba firmemente colocado a la *cabeza* de los tests de personalidad y se utilizaba en la consejería con sujetos normales, en ambientes laborales, médicos, militares y forenses así como con pacientes psiquiátricos. Para la década de los ochenta, la bibliografía sobre el MMPI incluía varias miles de referencias que documentaban, entre muchas otras cosas, el uso de sus 13 escalas básicas en una gran variedad de poblaciones, el desarrollo de centenares de escalas especiales basadas en sus reactivos y una vasta colección de correlatos empíricos de las escalas elevadas y modelos de perfil. Sin embargo, para ese momento, los problemas conceptuales documentados y psicométricos del inventario parecían aún más gravosos al considerar los avances en los campos de la psicopatología, la teoría de la personalidad y la elaboración de pruebas; para entonces ya se había demostrado también que las normas, anticuadas e insuficientes, resultaban inapropiadas para los examinados actuales, lo que implicaba que la base utilizada para determinar la anormalidad descansaba sobre fundamentos inciertos (Colligan, Osborne, Swenson y Offord, 1983, 1989). En efecto, la muestra original de estandarización se había convertido en algo similar a un grupo de referencia no normativo en términos del cual se definió la escala de calificación. Los datos mucho más extensos obtenidos más adelante, con referencia a las claves del perfil, proporcionaron la base para la interpretación normativa.

Así, la comisión encargada de la reestandarización del MMPI se enfrentó a la difícil tarea de modernizar el instrumento a la vez que trataba de salvar la riqueza interpretativa del material relevante para la evaluación de la personalidad, y en especial de la psicopatología, contenido en la estructura básica del MMPI. En un intento por mantener la continuidad, la comisión decidió conservar la mayor parte de los reactivos (todos los de las escalas clínicas y de validez originales y muchos de los reactivos de las escalas complementarias), lo que supone que se mantuvieron paralelamente sus debilidades inherentes. Los mayores cambios fueron una renormalización completa del inventario, el desarrollo de calificaciones T uniformes para las ocho escalas clínicas originales y para todas las escalas de contenido, la revisión y eliminación de los reactivos obsoletos o cuestionables, así como la inclusión de reactivos nuevos, la creación de nuevas escalas de validez (complementarias y de contenido) y la separación del inventario en dos versiones adecuadas para diferentes grupos de edad.

El Inventario Multifásico de Personalidad de Minnesota-2. Los reactivos del MMPI-2 consisten en 567 afirmaciones a las que el examinado da una respuesta de "cierto" o "falso". Salvo por algunos cambios en la redacción y un reordenamiento de los reactivos, los primeros 370 son prácticamente idénticos a los del MMPI, ya que sus respuestas se necesitaban para calificar las 10 escalas "clínicas" y las tres escalas de

[3] Dahlstrom y Dahlstrom (1980) reproducen los artículos originales.

"validez" originales. Los 197 reactivos restantes (107 de los cuales son nuevos) son necesarios para la calificación total de las 104 escalas y subescalas nuevas, revisadas, complementarias, de contenido y validez conservada que forman el inventario completo. El contenido de los reactivos varía considerablemente y cubre áreas como la salud general, afectiva, neurológica y los síntomas motores; las actitudes sexuales, políticas y sociales; aspectos educativos, ocupacionales, familiares y maritales; así como muchas manifestaciones bien conocidas de conducta neurótica o psicótica, como los estados obsesivo y compulsivo, los delirios, las alucinaciones, las ideas de referencia, las fobias y las tendencias sádicas y masoquistas. Dahlstrom (1993a) preparó un manual complementario que proporciona toda la información necesaria para comparar los reactivos del MMPI-2 con los originales. A continuación, se muestran algunos ilustrativos reactivos seguidos por el número que les corresponde en la forma actual de la prueba:[4]

Mi sueño es irregular y agitado. (39)
Creo que hay una conspiración en mi contra. (138)
Me preocupo por el sexo. (166)
Cuando me aburro me gusta crear cierta excitación. (169)
En el fondo a la mayoría de la gente le disgusta tener que ayudar a otras personas. (286)

El MMPI-2 proporciona calificaciones para 10 "escalas clínicas" básicas, que son las mismas del MMPI original:

1. Hs: Hipocondriasis
2. D: Depresión
3. Hi: Histeria
4. Dp: Desviación psicopática
5. Mf: Masculinidad-femineidad
6. Pa: Paranoia
7. Pt: Psicastenia
8. Es: Esquizofrenia
9. Ma: Manía
10. Is: Introversión social

Ocho de esas escalas se desarrollaron empíricamente en la década de los cuarenta por clave de criterio para los reactivos que diferenciaban pequeñas muestras clínicas (cada una formada por alrededor de 50 personas que representaban el tradicional diagnóstico psiquiátrico) y un grupo de control normal compuesto por 724 visitantes y familiares de los pacientes internos en los hospitales de la Universidad de Minnesota (Hathaway y McKinley, 1940, 1943). La escala de masculinidad-femineidad, que originalmente pretendía distinguir entre hombres homosexuales y heterosexuales, se desarrolló a partir de las diferencias en la frecuencia con que soldados varones y empleadas de una aerolínea apoyaban los reactivos. Las calificaciones en esta escala indican el grado al que los intereses y las actitudes del individuo coinciden con el estereotipo de su grupo sexual. La escala de introversión social, agregada después, se derivó de las respuestas de dos grupos contrastados de universitarios seleccionados sobre la base de sus calificaciones extremas en una prueba de introversión-extraversión.

[4] Tomado del Inventario Multifásico de Personalidad de Minnesota-2. Copyright © por The Regents of the University of Minnesota. Todos los derechos reservados. Reproducido con autorización

Al conservar intactas las escalas básicas del MMPI (salvo por la eliminación de nueve reactivos, el reordenamiento del resto y algunos cambios en la redacción), los autores del MMPI-2 buscaban preservar la riqueza de la información clínica asociada con la interpretación de las claves de los perfiles basados en los modelos de las puntuaciones obtenidas (Graham, 1993; Greene, 1991); conservaron las nociones obsoletas de psicopatología, las consecuencias de la ingenua e inadecuada aplicación del método empírico de grupos contrastados y otros problemas como la multidimensionalidad y el traslape de las escalas básicas (Helmes y Reddon, 1993).

El MMPI original se distingue por el uso de tres escalas llamadas de validez (conservadas en el MMPI-2),[5] que más que interesarse en los aspectos técnicos de la validez pretenden supervisar el descuido, la comprensión inadecuada, la simulación y la existencia de ciertas actitudes y predisposiciones de respuesta. Las calificaciones de validez comprenden:

Puntuación de mentiras (L). Se basa en un grupo de reactivos que hacen que el examinado aparezca bajo una luz favorable pero que es poco probable que se respondan honestamente en la dirección favorable (por ejemplo: no me gusta nadie que conozco).

Puntuación de infrecuencia (F). Determinada por un conjunto de 60 (de un original de 64) reactivos en la dirección calificada por no más del 10 por ciento del grupo de estandarización del MMPI. Aunque representan conductas indeseables, esos reactivos no corresponden a ningún patrón particular de anormalidad, por lo que es poco probable que haya alguien que muestre todos o casi todos los síntomas. Una alta puntuación en F puede indicar errores de calificación, descuido al responder, una gran excentricidad, procesos psicóticos o simulación deliberada.

Puntuación de corrección (K). A partir de otra combinación de reactivos especialmente elegidos, esta puntuación ofrece una medida de las actitudes de evaluación que se consideran más sutiles. Una alta puntuación en K puede indicar que se está a la defensiva o un intento deliberado de "simular que se está mal".

Las puntuaciones de las dos primeras escalas (L y F) suelen utilizarse para hacer una evaluación global del registro de la prueba, que se considera invalidado si cualquiera de ellas excede del valor especificado. Por otro lado, la puntuación de la escala K funciona como variable de supresión, se utiliza para calcular un factor de corrección que se agrega a las puntuaciones de algunas escalas clínicas a fin de obtener totales ajustados. La eficacia de este uso de la puntuación K es cuestionable, por lo que las puntuaciones de las escalas afectadas pueden informarse con y sin la corrección. Aunque una puntuación K inusualmente alta podría hacer que un registro resulte sospechoso e indicar la necesidad de un escrutinio adicional, las elevaciones moderadas reflejan la fortaleza del ego y un ajuste positivo, esto es importante si se considera la historia y las circunstancias de la vida del individuo al evaluar estas elevaciones.

Entre las 21 escalas complementarias del MMPI-2, se encuentran los tres nuevos indicadores de "validez" que pueden ayudar a evaluar el cuidado y la veracidad con

[5] También hay una puntuación de "No lo sé (?)" que representa el número de reactivos en el que se hicieron dos marcas o se omitieron. Si este conteo excede a los 30 reactivos, el registro de la prueba se considera sospechoso y probablemente no válido.

que los examinados contestan el inventario. Éstas son la escala F posterior (F_p), la escala de inconsistencia de respuestas variables (INVAR) y la escala de inconsistencia de respuestas verdaderas (INVER). Mientras que la escala F_p es básicamente una extensión de la escala F original para los reactivos que aparecen en la segunda mitad del inventario, las escalas INVAR e INVER son nuevas y constan de pares de reactivos con similares significados opuestos que pretenden detectar las respuestas inconsistentes o contradictorias.

La forma básica del perfil para el MMPI-2 (figura 13.1) abarca las 13 escalas clínicas y de validez conservadas de la versión original. También, cuenta con perfiles separados para las 15 escalas de contenido, las 27 del componente de contenido, las 21 complementarias y las 28 subescalas Harris-Lingoes.[6] Algunas de las escalas y subescalas son nuevas y otras se conservaron del original, pero todas se califican utilizando la muestra normativa del MMPI-2 compuesta por 2 600 adultos de 16 a 84 años. Ésta se extrajo de siete estados en un intento por reflejar a la población norteamericana en términos de variables demográficas significativas como el género, la edad y la etnicidad, por lo que resulta mucho más representativa de la población actual de los Estados Unidos que el grupo normativo original de Minnesota (Dahlstrom y Tellegen, 1993). Pese a ello, se ha cuestionado la representatividad de la muestra debido principalmente a los elevados niveles ocupacionales y de escolaridad, y por la poca representación de hispanos y asiaticoamericanos en comparación con las cifras del censo de 1980 (Duckworth, 1991).

El resultado principal del establecimiento de nuevas normas para el MMPI fue una disminución en la elevación de las puntuaciones de los perfiles clínicos. Este cambio, que ya se anticipaba, obedece quizá a diferencias generacionales así como a factores especiales que son exclusivos de la muestra original de Minnesota y a la forma en que se utilizó en la elaboración del MMPI (D. S. Nichols, 1992). En cualquier caso, el corte de la calificación T necesaria para considerar que una elevación de la escala es de interés clínico ya no es de 70 sino de 65, o aproximadamente 1.5 *DE* por encima de la media. Otra innovación introducida en el MMPI-2 es el uso de calificaciones T uniformes, en contraste con las normalizadas o derivadas linealmente en ocho de las 10 escalas clínicas y en todas las de contenido. Esto implica igualar las calificaciones en todas esas escalas a una distribución promedio compuesta. Las calificaciones T uniformes permiten hacer comparaciones entre escalas en términos de equivalentes percentilares sin distorsionar de manera notable la forma de las distribuciones de las puntuaciones crudas, las cuales son positivamente asimétricas, aunque en diferentes grados (Tellegen y Ben-Porath, 1992).

Los cambios descritos se justificaron sobre bases psicométricas; sin embargo, se generó una controversia sobre la posibilidad de generalizar para el MMPI-2 el conocimiento clínico y los descubrimientos en la bibliografía sobre la interpretación del perfil del MMPI, y esto porque los cambios mencionados produjeron diferencias entre los patrones de los perfiles y las claves obtenidas del MMPI y del MMPI-2 (Chojnacki y Walsh, 1992; Morrison, Edwards y Weissman, 1994; Tellegen, Ben-Porath, 1993). Los datos sugieren que para los perfiles bien definidos con una clara separación entre

[6] Como son de 1996, el editor del MMPI-2 todavía no pone en circulación las 10 subescalas de sutil-obvio de Wiener-Harmon.

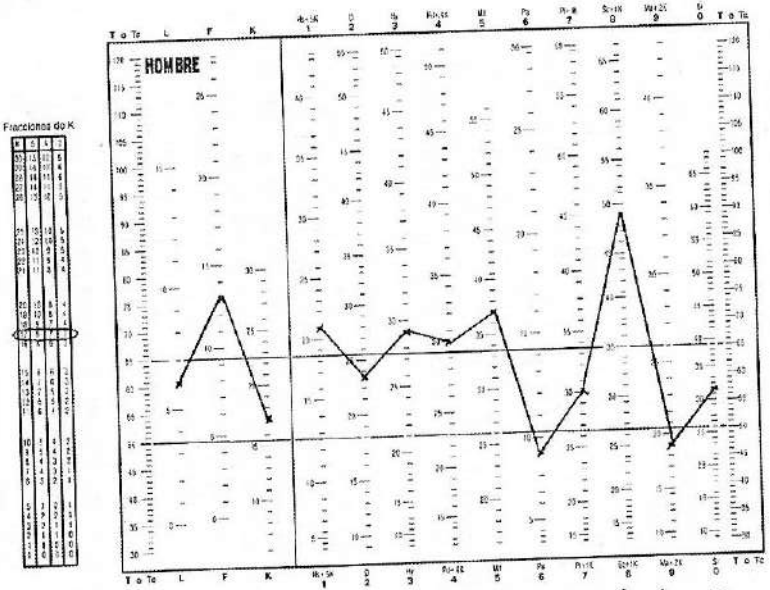
MMPI-2 S.R. Hathaway and J.C. McKinley *Minnesota Multiphasic Personality Inventory-2*

Perfil para escalas básicas

Minnesota Multiphasic Personality Inventory-2
Copyright © by THE REGENTS OF THE UNIVERSITY OF MINNESOTA
1942, 1943 (renewed 1970), 1989. This Profile Form 1989.
All rights reserved. Distributed exclusively by NATIONAL COMPUTER SYSTEMS, INC.
under license from The University of Minnesota.

"MMPI-2" y el inventario multifásico de personalidad de Minnesota-2 son marcas registradas propiedad de la Universidad de Minnesota. Impreso en los Estados Unidos.

Nombre ID # 511
Dirección
Ocupación — Fecha de evaluación / /
Educación 1943. Edad 41 Estado civil
Referido por
Clave del MMPI-2 8"51'34+2-70/96: F'+L-K/
Iniciales del examinador



Puntuación cruda 6 13 17 12 23 29 23 37 9 13 32 16 32
? Puntuación cruda 7
K que debe agregarse 9 7 17 17 3
Puntuación cruda con K 21 30 30 49 19

NATIONAL COMPUTER SYSTEMS NCS

24001

Figura 13.1. Perfil de las escalas básicas del MMPI-2 de un hombre de 41 años. La clave resume las relaciones entre las calificaciones en las escalas del MMPI-2 y proporciona una base para la interpretación y la composición de perfiles.

(Las puntuaciones en la ilustración del perfil se tomaron de los *Annotated Sample Report for The Minnesota Report Adult Clinical System-Revised* [Universidad de Minnesota, 1993, pp. 2-7].)

las calificaciones de la escala parece haber tanta consistencia entre los tipos de códigos del MMPI y del MMPI-2 como la que se encuentra al repetir la aplicación de cualquiera de esas versiones (Archer, 1992b; Graham, 1993). En cualquier caso, el manual del MMPI-2 contiene información que permite al usuario comparar las calificaciones generadas de las dos versiones, basándose en las respuestas a cualquiera de ellas, y si bien esta aproximación no está exenta de problemas (véase, por ejemplo, Ben-Porath y Tellegen, 1995), algunos consideran que es un método empíricamente defendible que permite a los usuarios superar el periodo de transición entre las dos versiones (véase, por ejemplo, Humphrey y Dahlstrom, 1995).

El Inventario Multifásico de Personalidad de Minnesota para Adolescentes. El MMPI-A es la nueva forma del inventario diseñada específicamente para adolescentes, aunque incorpora casi todos los rasgos del MMPI y del MMPI-2 (incluyendo las 13 escalas básicas). Para adaptar la prueba a los más jóvenes se redujo la longitud del inventario a 478 reactivos, se incluyeron nuevos reactivos y escalas que cubren áreas específicamente relevantes (como la escuela y los problemas familiares) y, sobre todo, se proveyeron normas apropiadas para esa edad. El MMPI-A utilizó una muestra normativa de 1 620 adolescentes contemporáneos de 14 a 18 años. Al mismo tiempo, se obtuvo una muestra clínica de 713 adolescentes con el mismo rango de edad para comparaciones y estudios de validez.

Además de las escalas clínicas y de validez que comparte con el MMPI-2, el MMPI-A tiene sus propias escalas de validez (F1 y F2), así como algunas escalas y subescalas complementarias y de contenido exclusivas y otras que son comunes a ambos instrumentos. Aunque numerosas investigaciones, incluidas las normas y tablas de conversión publicadas por Dahlstrom *et al.* (1972) y Marks, Seeman y Haller (1974), apoyan el uso del MMPI en adolescentes, esta investigación no necesariamente se aplica al MMPI-A, que más que una revisión es un instrumento nuevo, por lo que su utilidad deberá ser determinada por las investigaciones y los materiales interpretativos que se iniciaron con su publicación (Archer, 1992a; Butcher y Williams, 1992; Williams, Butcher, Ben-Porath y Graham, 1992).

Comentarios finales sobre los Inventarios Multifásicos de Personalidad de Minnesota. A pesar de su origen como prototipo de un empirismo aplicado ingenuamente y de los rumores constantes sobre su inminente desaparición, el MMPI se las ha arreglado para sobrevivir. De hecho, aunque su propósito original era ayudar en el proceso de clasificación psiquiátrica y aunque los procedimientos seguidos en su desarrollo lo hacían inapropiado para la evaluación de la personalidad en individuos normales, se ha utilizado lo mismo con personas normales que con pacientes psiquiátricos. El instrumento y sus versiones revisadas han añadido una serie de procedimientos y estrategias interpretativas al marco original, empíricamente derivado de la estructura, de los inventarios. Dichos rasgos, que aunque ya se han mencionado merecen una mayor consideración, incluyen muchas de las escalas desarrolladas al agrupar los reactivos sobre la base de su contenido (Butcher, Graham, Williams y Ben-Porath, 1990) así como el uso del análisis factorial en el desarrollo de algunas de sus escalas complementarias (Welsh, 1956).

Continúan evolucionando nuevas formas de aproximarse a la compleja tarea de interpretación del MMPI. Una de las más recientes es el uso de compendios estructurales para dar cierta coherencia y facilitar el uso de la multiplicidad de escalas correlacionadas producidas por los inventarios de Minnesota. El método, que se basa en el análisis de las escalas, subescalas y reactivos, pretende reducir el número de dimensiones necesarias para interpretar los resultados de los inventarios por corte transversal de clasificaciones arbitrarias (como las escalas "suplementaria", "de contenido" y de "Harris y Lingoes"). Las dimensiones más destacadas se utilizan para organizar las categorías en un formato "resumido estructural" similar al desarrollado por Exner para el Rorschach (véase el capítulo 15). Esta aproximación a la interpretación del MMPI se encuentra en una etapa incipiente de desarrollo y requiere de mayor investigación y validación; no obstante, ya se aplica de una manera que parece prometedora tanto para el MMPI-2, en el que las dimensiones estructurales se derivaron sobre todo a partir del análisis del contenido, como para el MMPI-A, cuyas dimensiones se derivaron mediante el análisis factorial (Archer y Krishnamurthy, 1994; Archer, Krishnamurthy y Jacobson, 1994; Nichols y Greene, 1995).

Otros dos desarrollos que siguen los pasos del MMPI-2 y el MMPI-A, al igual que la mayor parte de las pruebas, son la automatización de procedimientos para la aplicación, calificación e interpretación de los inventarios y las traducciones de los instrumentos a otros idiomas. Al igual que con el MMPI, se dispone para los nuevos inventarios de procedimientos convencionales de aplicación y calificación computarizada, al igual que varias formas automatizadas de interpretación. Además, se ha elaborado y aplicado con resultados prometedores una versión adaptada y computarizada del MMPI-2 (Roper, Ben-Porath y Butcher, 1991, 1995).

Aunque pasó casi una década para que el MMPI original fuera traducido por primera vez, el trabajo sobre las adaptaciones transculturales del MMPI-2 empezó incluso antes de que éstas se publicaran. Durante los tres primeros años de su existencia, había 15 proyectos de traducción del MMPI-2 completos o en proceso. Se prepararon seis traducciones o adaptaciones al español, incluidas dos versiones para uso con hispanos en los Estados Unidos, y también se dispone de un manual sobre traducciones y adaptaciones internacionales del MMPI-2 (Butcher, 1996). Un descubrimiento digno de mención de quienes han recopilado los datos para el uso de dichas traducciones y adaptaciones es que las muestras transculturales normales tienen una calificación más cercana a las normas de estandarización del MMPI-2 comparadas con las normas de las muestras anteriores del MMPI.

Al tratar de mejorar el ejemplo clásico del inventario de personalidad sin alterarlo de manera fundamental, la Comisión de Reestandarización del MMPI se impuso dos metas difíciles y, en gran medida, contradictorias. El tiempo dirá si las decisiones que ha tomado han de extender la supremacía de los inventarios de Minnesota al próximo siglo o si serán superados por una nueva generación de instrumentos similares, como el Inventario Básico de Personalidad de Jackson (*Basic Personality Inventory*, que estudiamos en otra sección de este capítulo) o el Inventario de Evaluación de la Personalidad (*Personality Assessment Inventory*, PAI) desarrollado por Leslie Morey (1991) que utiliza una elaborada estrategia secuencial que combina métodos lógicos y empíricos para garantizar la firmeza psicométrica de sus escalas. Mientras tanto, continúa la di-

versidad y amplitud de publicaciones de libros y artículos sobre el MMPI-2 y el MMPI-A (Butcher, 1990; Butcher, Graham y Ben-Porath, 1995; Keller y Butcher, 1991; Pope, Butcher y Seelen, 1993).

Inventario Psicológico de California

Desde su publicación, el MMPI ha servido como base para la preparación de otros inventarios de gran difusión. Un ejemplo notable es el Inventario Psicológico de California (*California Psychological Inventory*, CPI), que, si bien derivó casi la mitad de sus reactivos del MMPI, se desarrolló para usarse con poblaciones adultas normales. La tercera edición del CPI consta de 434 reactivos que deben contestarse con "cierto" o "falso" y produce puntuaciones en 20 escalas (Gough y Bradley, 1996). Tres de ellas, que son escalas de "validez" diseñadas para evaluar las actitudes del examinado, se conocen como: escala de Bienestar (Be), que se basa en las respuestas de personas normales a las que se pide "simular para causar una mala impresión"; escala de Buena impresión (Bi), basada en las respuestas de personas normales a las que se pide "simular para causar una buena impresión", y escala de Comunalidad (Cm), que se basa en el conteo de la frecuencia de las respuestas más populares. Las 17 escalas restantes proporcionan calificaciones en dimensiones de personalidad como Dominancia, Sociabilidad, Autoaceptación, Responsabilidad, Socialización, Autocontrol, Logro mediante la conformidad, Logro mediante la independencia, Empatía e Independencia. Las dos últimas escalas se agregaron en la revisión de 1987.

En 13 de esas 17 escalas, los reactivos se seleccionaron sobre la base de las respuestas de grupos contrastados, y la obtención de criterios como las calificaciones escolares, la pertenencia a una clase social, la participación en actividades extracurriculares y las valoraciones. Estas últimas se obtuvieron mediante nominaciones de los pares, que constituyen una técnica eficaz de evaluación de muchos rasgos interpersonales (véase el capítulo 16). Los reactivos de las cuatro escalas restantes originalmente se agruparon de manera subjetiva y luego se comprobó su consistencia interna. La validación cruzada de todas las escalas con muestras de tamaño considerable ha producido diferencias grupales significativas, si bien el traslape de los grupos de criterio contrastados es considerable y las correlaciones con el criterio a menudo son bajas.

Como en el MMPI-2, todas las calificaciones del CPI se informan en términos de una escala de calificación estándar con una media de 50 y una *DE* igual a 10; en la actualidad, la escala se basa en una muestra no normativa de 3 000 mujeres y 3 000 hombres cuyos datos se extrajeron de los archivos del CPI para representar la población general de los Estados Unidos en términos de edad, nivel socioeconómico y área geográfica. Se proporcionan normas por separado para mujeres, para hombres y para ambos géneros combinados, y se ofrecen medias y *DE* de las calificaciones de cada escala para muchos grupos especiales.

El CPI, publicado originalmente en 1956, se pensó como un "sistema abierto" del que podían eliminarse o agregarse elementos según se necesitara (Gough, 1987, p. 1). Así, en las sucesivas revisiones, el inventario acortó su longitud original (de 480 reactivos a 462 en la revisión de 1987 y más recientemente a 434) para eliminar los reactivos que podrían ser objetables para algunos de los examinados o legalmente cuestionables de acuerdo con la Ley de Estadounidenses con Discapacidades de 1990 (P. L., 101–336),

especialmente en la selección y colocación de personal. Al utilizar el extenso archivo de datos de más de 13 000 sujetos en todos los reactivos del CPI, Gough y Bradley se han esforzado por conservar la confiabilidad y validez de las escalas básicas, manteniendo constante su longitud mediante la sustitución de los reactivos eliminados con otros funcionalmente equivalentes en términos del grado de correlación con los criterios de pertenencia a la escala.

La investigación con el CPI ha proporcionado información de gran utilidad para analizar los perfiles basados en elevaciones de la escala individual y las configuraciones o los patrones de puntuaciones en dos o más escalas, de una manera afín a la tradición de interpretación de los códigos del MMPI (McAllister, 1996). Los estudios transculturales indican que el CPI es útil para explorar las diferencias de personalidad entre grupos étnicos (véase Dana, 1993; Davis, Hoffman y Nelson, 1990). Además de las escalas básicas se han elaborado otras, como la de Potencial Ejecutivo y una de Orientación al Trabajo (Gough, 1984, 1985), y guías de aplicación que hacen especial referencia al uso del CPI en la selección de personal y el desarrollo profesional (Meyer y Davis, 1992).

Desde su revisión en 1987, el CPI también incluye un modelo tipológico tridimensional para la clasificación de individuos con altas y bajas puntuaciones en las tres escalas estructurales o vectores identificados por los análisis factoriales y de reactivos. Las escalas estructurales miden las dimensiones de orden superior de Internalidad contra Externalidad, Aceptación contra Rechazo de las Normas y Autorrealización. Las calificaciones en las dos primeras escalas se utilizan para categorizar a los examinados en cuatro tipos de personalidad (Alfa, Beta, Gamma y Delta), mientras que las calificaciones de la tercera escala vector pretenden medir el nivel de integración o realización del potencial positivo asociado con el tipo que corresponda al individuo. El modelo tipológico es atractivo para los usuarios organizacionales del inventario, pero ha sido criticado por la vaga descripción de los procedimientos empleados para derivarlo (Engelhard, 1992). Además, comparte las debilidades conceptuales y psicométricas inherentes a la clasificación de la gente por tipos basada en la dicotomización arbitraria de dos o más dimensiones continuas.[7]

Inventarios de personalidad para niños

Aunque no utilizan los reactivos o datos del MMPI, el Inventario de Personalidad para Niños (*Personality Inventory for Children*, PIC) se elaboró con la misma metodología general que el MMPI y el CPI (Wirt y Lachar, 1981; Wirt, Lachar, Klinedinst y Seat, 1991). Destinado para niños y adolescentes de tres a 16 años, el PIC se desarrolló en la Universidad de Minnesota luego de unos 20 años de estudios conducidos por un grupo de investigadores que habían recibido una concienzuda formación en la lógica y el uso clínico del MMPI. Una diferencia importante entre el PIC y el MMPI atañe a la forma en que se obtiene la información: a los reactivos de cierto o falso no da respuesta el niño, sino un adulto que lo conoce bien, por lo general la madre. El procedimiento es congruente con la práctica común en las clínicas infantiles de entrevistar a los padres como la fuente principal de información sobre los problemas que presenta el niño y la historia del caso. En efecto, este inventario proporciona una for-

[7] Véase el análisis del Indicador tipo Myers-Briggs en el capítulo 16.

ma sistemática de obtener dicha información y de interpretarla en términos de los datos normativos y diagnósticos.

El PIC original comprendía un total de 600 reactivos agrupados en tres "escalas de validez", una de detección general y 12 escalas clínicas. Las de validez incluyen la escala de Mentiras, que consta de reactivos que hacen que el niño aparezca de una manera irrealmente favorable; la escala de Frecuencia, que comprende reactivos rara vez apoyados, y la escala Defensiva, diseñada para evaluar la tendencia de los padres a ponerse a la defensiva respecto a la conducta del niño. La escala de detección, llamada de Ajuste, se utiliza para identificar a los niños que requieren de evaluación psicológica. Las 12 escalas clínicas se diseñaron para evaluar el desarrollo cognoscitivo y el aprovechamiento académico del niño, varias formas bien establecidas de problemas emocionales e interpersonales (por ejemplo, depresión, ansiedad, aislamiento e hiperactividad), así como el clima psicológico de la familia.

En la versión actual del Formato Revisado del PIC (PIC-R), los reactivos se reordenaron y se redujeron a 420; el cuadernillo de aplicación los agrupa en tres partes progresivamente más largas. La Parte I (reactivos 1–131) produce puntuaciones en la escala de Mentiras y para cuatro nuevas escalas basadas en factores más amplios. La Parte II (reactivos 132–280) *agrega* a la Parte I versiones abreviadas de otras escalas regulares y una lista parcial de reactivos críticos. La Parte III (reactivos 281–420) añade el resto de los reactivos requeridos para el complemento de las 16 calificaciones de la escala original, cuatro calificaciones de factor amplio y el conjunto completo de reactivos críticos.

Siete de las 16 escalas PIC originales se desarrollaron mediante la comparación empírica de las frecuencias de respuesta en grupos de criterio y control. Gracias a un procedimiento iterativo se fueron agregando los reactivos hasta que la escala alcanzó una validez óptima. En las nueve escalas restantes se siguieron procedimientos de validación de contenido en los que los reactivos se seleccionaban inicialmente sobre la base de las elecciones de los jueces o las valoraciones de la relevancia de la escala; sin embargo, incluso en esos casos la evaluación de la consistencia interna de las respuestas a los reactivos en las escalas y el análisis factorial de los reactivos contribuyeron a la validación de constructo de las escalas.

Una monografía preparada por Lachar y Gdowski (1979) ofrece una buena cantidad de datos interpretativos para el PIC original que se basan en un estudio amplio y sistemático de validación. El manual complementario preparado para la versión del Formato Revisado hace lo mismo para las escalas de factores y proporciona los datos psicométricos sobre las escalas abreviadas (Lachar, 1982). Además, Lachar y sus asociados realizaron análisis de agrupamientos para clasificar muestras heterogéneas de niños sobre la base de sus perfiles PIC y estudiaron las características de valor diagnóstico significativo asociadas con diferentes perfiles PIC. Tomando como base esta línea de investigación, se han implantado reglas secuenciales para la clasificación de perfiles en tipos y procedimientos para calcular índices de similitud de perfiles (Gdowski, Lachar y Kline, 1985; Kline, Lachar y Gdowski, 1992). Esta aproximación actuarial a la interpretación, que también se ha utilizado con el MMPI y el CPI, es una extensión y elaboración de la tradición de clave empírica de criterio para la fase de desarrollo de una prueba a su fase interpretativa (véase, por ejemplo, Kline, Lachar y Boersma, 1993).

Es importante aclarar que el PIC-R no es un inventario autodescriptivo, sino de conducta observada, por lo que resulta compatible con la orientación conductual de la

evaluación a la psicología clínica que explicaremos en el capítulo 17. Sin embargo, los informes de los padres tienen limitaciones, que en general reconocen y advierten los autores del PIC, quienes señalan que las respuestas pueden reflejar en parte su motivación, sus actitudes y normas personales o culturales. Por esto, se espera encontrar, como efectivamente sucede, ciertas inconsistencias en los informes de observadores distintos (ambos padres) y entre los informes de los padres y las autodescripciones de los niños. Una forma de manejar este problema inevitable consiste en evaluar las predisposiciones especiales de respuesta que pueden haber distorsionado los datos mediante las "escalas de validez". La segunda opción consiste en obtener y comparar los informes de más de un observador. La tercera vía consiste en obtener y comparar datos de la autodescripción y del observador.

Para tal fin se ha desarrollado el Inventario de Personalidad para Jóvenes (*Personality Inventory for Youth*, PIY) como medida autodescriptiva paralela que se equipara al PIC-R. Aunque el PIY puede utilizarse solo, idealmente proporciona la base para una evaluación más completa, usando en conjunto los perfiles de padres e hijos (Lachar y Gruber, 1995a, 1995b). Los autores del PIY se sirvieron de los primeros 280 reactivos del PIC-R como punto de partida para la formación del banco de reactivos, de los que la mayor parte se cambió de la tercera a la primera persona con algunas modificaciones menores (por ejemplo, "Mi hijo a menudo trae amigos a casa" se convirtió en "A menudo llevo amigos a casa"). Hubo que hacer mayores correcciones a algunos para adaptarlos al significado pretendido; otros tuvieron que descartarse por ser inapropiados para la autodescripción para el nivel de edad objetivo (se estableció de los nueve a los 18 años) debido a los problemas que supone la autodescripción en las edades más tempranas. Algunos reactivos son nuevos, de modo que la versión final del PIY consta de 270 ítems que comprenden nueve escalas clínicas que no se traslapan, 24 subescalas no traslapadas y cuatro "escalas de validez". Los primeros 80 reactivos pueden usarse como forma abreviada para propósitos de identificación. El PIY se desarrolló y estandarizó utilizando una muestra de 2 337 estudiantes en grupos de educación regular y una muestra clínica de 1 178 niños y adolescentes (Lachar y Gruber, 1993).

Aún queda mucho trabajo por hacer con el PIC-R y el PIY. Es necesario actualizar las normas del PIC (obtenidas a finales de los cincuenta y principios de los sesenta) y extender su base de investigación a los preescolares (Knoff, 1989). Hasta donde concierne al PIY, todavía queda por establecer su utilidad clínica solo y en combinación con el PIC-R; sin embargo, ambos instrumentos comparten una base empírica impresionante y la ventaja de proporcionar un conjunto integrado de instrumentos multidimensionales específicamente diseñados para usarse con niños y adolescentes.

EL ANÁLISIS FACTORIAL EN EL DESARROLLO DE PRUEBAS

En un esfuerzo por llegar a una clasificación sistemática de los rasgos de personalidad, algunos psicólogos recurrieron al análisis factorial, una técnica (ya estudiada en relación con la organización de las habilidades cognoscitivas) que es ideal para la tarea de reducir el número de categorías necesarias para explicar los fenómenos conductuales al buscar patrones constantes en su ocurrencia. Se recordará que, dado que el proceso depende de un examen de las pruebas o los reactivos que tienen la mayor carga en

cada factor, es probable que un elemento de subjetividad entre en la identificación de los factores (véase el capítulo 11). De ahí que la identificación cruzada de factores de investigaciones separadas que utilizan mediciones distintas resulte difícil y haya sido fuente de discrepancias en cuanto al número de rasgos y a los títulos que se les asignan. Además, en el estudio de los rasgos de personalidad, el análisis factorial puede aplicarse de muchas formas distintas, de las cuales durante varias décadas han coexistido dos grandes tradiciones de investigación que, hasta cierto punto, se han unificado en los años recientes.

La serie de estudios realizada por Guilford y sus colaboradores (véase, Guilford, 1959, capítulo 16; Guilford y Zimmerman, 1956) ejemplifica una de esas tradiciones, centrada en el uso de los datos de cuestionarios de personalidad. En lugar de correlacionar las puntuaciones totales de los inventarios, los investigadores calcularon las correlaciones entre los reactivos individuales de muchos inventarios de personalidad. Como subproducto de su investigación se desarrollaron tres inventarios de personalidad que al final se combinaron en el Estudio de Guilford y Zimmerman sobre el Temperamento, un inventario que produce puntuaciones separadas para 10 rasgos (por ejemplo, Restricción, Dominio, Estabilidad Emocional y Cordialidad), cada puntuación basada en 30 reactivos diferentes.

Otra de las primeras aplicaciones de los métodos factoriales al desarrollo de inventarios de personalidad, en ocasiones denominada como la tradición "léxica", se encuentra en el trabajo que inició R. B. Cattell en la década de los cuarenta (John, Angleitner y Ostendorf, 1988). Para tratar de obtener una descripción cabal de la personalidad, Cattell comenzó por recopilar todos los nombres de rasgos de personalidad que aparecían en el diccionario (como lo hicieron Allport y Odbert en 1936) o en la bibliografía psiquiátrica y psicológica. Esta lista de casi 18 000 términos se redujo inicialmente al reunir los sinónimos evidentes, y la lista más pequeña se utilizó para obtener las estimaciones asociadas de un grupo heterogéneo de adultos. Las correlaciones y los análisis factoriales de estas estimaciones y de los datos obtenidos en cuestionarios autodescriptivos permitieron la identificación de lo que Cattell describió como "la fuente de los rasgos primarios de personalidad", una designación que parece implicar más universalidad y estabilidad en los resultados de lo que parece justificado por la investigación antecedente. Una característica de la aproximación de Cattell es que no considera al análisis factorial como una técnica de reducción de datos, sino como un método para descubrir los rasgos causales subyacentes (Cattell, 1979).

El Cuestionario de Dieciséis Factores de Personalidad de Cattell (16PF). Tomando como base su investigación factorial, Cattell y sus colaboradores desarrollaron una serie de inventarios de personalidad, de los cuales el más conocido es el Cuestionario de Dieciséis Factores de Personalidad, actualmente en su quinta edición (Cattell, Cattell y Cattell, 1993; Conn y Rieke, 1994; Russell y Karol, 1994). Publicado originalmente en 1949, el 16 PF se diseñó para edades que rebasan los 16 años y produce 16 puntuaciones en rasgos como Audacia Social, Dominancia, Vigilancia, Estabilidad Emocional y Conciencia de las Reglas. Al paso de los años, los 16 factores, identificados por las mismas letras en las distintas ediciones del cuestionario, se han perfeccionado y renombrado, como sucedió al descartar la terminología esotérica que

Cattell empleó originalmente para designar los rasgos; por ejemplo, los extremos que soporta la dimensión ahora conocida como Audacia Social inicialmente se conocían como "Threctia" y "Parmia", en los extremos tímido y atrevido, respectivamente. Los factores de segundo orden del 16 PF, que previamente fluctuaban entre cuatro y ocho, ahora se conocen como "factores globales" y se limitan a cinco, número que equipara al popular modelo de cinco factores que veremos en la siguiente sección.

En la quinta edición sólo se dispone de una forma del 16 PF que contiene 185 reactivos, que en su mayor parte se seleccionaron de las formas previas del cuestionario en función de su contenido y de la calidad de sus propiedades psicométricas. Se obtuvieron nuevas normas para el 16 PF en una muestra de 2 500 individuos seleccionados para representar aproximadamente a la población estadounidense del censo de 1990 en términos de género, raza, distribución por edad y educación. Una de las características distintivas del 16 PF es la inclusión, bajo el rubro de "Preguntas para la Resolución de Problemas", de 15 reactivos que se presentan al final del inventario e incluyen la escala de Razonamiento que pretende hacer una medición rápida de la capacidad mental. Además, el cuestionario cuenta ahora con tres índices de estilos de respuesta diseñados para evaluar la aquiescencia, las respuestas aleatorias y los intentos del individuo por presentarse de una manera poco realista, como si tuviera cualidades socialmente deseables o indeseables.

En comparación con las ediciones anteriores, en la quinta hay una mejora en la consistencia interna y la confiabilidad test-retest de las escalas de los 16 factores primarios del 16 PF. Además, el manual técnico de la edición contiene mayor información sobre la validez que los manuales anteriores. Ahora bien, parece persistir el problema de la falta de independencia factorial de las 16 escalas primarias, que ya era evidente en las ediciones previas. Este problema se ha acentuado por el hecho de que otros investigadores que emplean las variables originales de Cattell no han logrado repetir algo cercano a los 16 factores, sino apenas entre cuatro y siete (Digman, 1990; L. R. Goldberg, 1993). D. W. Fiske (1949) realizó los primeros intentos por repetir los factores primarios de Cattell y llegó a una solución de cinco. Su trabajo se cita ahora como la primera versión del modelo contemporáneo que estudiaremos a continuación.

El "modelo de los cinco factores" y por qué funciona. Los trabajos actuales sobre la evaluación de la personalidad conceden gran atención al llamado Modelo de los Cinco Factores, que representa un nivel inusual de consenso entre los investigadores de la personalidad de diversas tradiciones del análisis factorial (Costa y Widiger, 1994; Digman, 1990; McCrae y John, 1992; Wiggins y Pincus, 1992). Al mismo tiempo, la forma particular en que se presenta el modelo ha generado grandes críticas y controversias (Block, 1995; Carlson, 1992; Goldberg, 1993; Kroger y Wood, 1993; Loevinger, 1994).[8] En esencia, el modelo pretende utilizar un patrón jerárquico de análisis para simplificar la enorme cantidad de datos sobre la conducta afectiva de los individuos, lo que hace más fácil manejar la información al evaluar a individuos y al predecir su comportamiento en situaciones determinadas. Los factores son más descriptivos que explicativos, y no son más básicos que los reactivos individuales o las pruebas de las que se derivaron.

[8] Véase también otros artículos que siguen a Kroger y Wood en la misma sección de Comentarios del *American Psychologist*, 1993, pp. 1298-1304.

El Modelo de los Cinco Factores es similar a la estructura jerárquica obtenida del análisis factorial de los tests de habilidad que estudiamos en el capítulo 11. Aunque en las últimas cuatro décadas varios investigadores han coincidido en que se necesitan cinco factores para explicar la mayor parte de las correlaciones entre una gran cantidad de datos descriptivos de la personalidad, el número apropiado de factores para propósitos particulares puede caracterizarse mejor como 5 ± 2. Por otro lado, también hay controversia en relación con los nombres dados a cada factor (véase, por ejemplo, Digman, 1990, p. 423; Loevinger, 1994; Paunonen, 1993), lo que es de esperar, toda vez que los factores reflejan diferentes selecciones de inventarios, escalas, formatos de respuesta y muestras de las que se derivaron.

La forma en que se presentó el modelo de los cinco factores también generó cierta confusión y malos entendidos, pues implicaba que se habían "descubierto" los factores básicos de la personalidad mediante alguna nueva aproximación. En realidad, los factores corresponden al segundo nivel de una jerarquía que resulta del análisis factorial de los datos de tests de personalidad y escalas de valoración. Es importante advertir que el segundo nivel de la jerarquía, que produce factores de amplitud media, ha demostrado ser más práctico y fácil de replicar, tanto en la evaluación cognoscitiva como en la afectiva. Esos factores producen puntuaciones más confiables, de ahí que la validez resultante sea, en correspondencia, más alta porque agregan o combinan medidas de los niveles inferiores de descriptores conductuales y rasgos más estrechamente definidos. Si una batería de pruebas se elabora de acuerdo con un modelo jerárquico, como en las Escalas de Habilidad Diferencial que vimos en el capítulo 8, la batería puede aplicarse con enorme flexibilidad para adaptarse a propósitos específicos; por ejemplo, después de identificar el factor de nivel medio que mejor caracteriza al individuo, el análisis puede complementarse al examinar sus puntuaciones en las pruebas más detalladas del nivel inferior.

Los dos investigadores más vinculados al modelo de los cinco factores prepararon una prueba que se ajusta a su versión del modelo. El Inventario de Personalidad NEO, Revisado (*Revised NEO Personality Inventory*,[9] NEO PI-R, Costa y McCrae, 1992b) proporciona en su edición actual puntuaciones para las cinco dimensiones o *dominios* principales de la personalidad y para 30 rasgos o *facetas* adicionales que identifican a cada dominio. Costa y McCrae evitan el uso del término "factores" para designar las escalas o componentes que pueden aplicarse a cualquier nivel de la jerarquía. En la tabla 13.1 se presentan los cinco dominios principales —Neuroticismo (N), Extraversión (E), Apertura a la Experiencia (O), Agradabilidad (A) y Escrupulosidad (C)— y sus respectivas facetas.

Las escalas del NEO PI-R se elaboraron luego de 15 años de investigación que comenzó con una serie de estudios longitudinales del envejecimiento en muestras de adultos normales y que luego se extendió a muestras clínicas, laborales y universitarias. Aunque se designó como medida de "rasgos de la personalidad normal", Costa y McCrae pretenden que el instrumento sea tan útil en la clínica y otros medios aplicados como en la investigación. Entre las innovaciones metodológicas introducidas en el NEO PI-R está el acceso a una forma autodescriptiva (forma S) y dos versiones de

[9] Las siglas NEO representan a Neuroticismo, Extraversión y Apertura a la Experiencia, pero las palabras no se utilizan en el título completo del inventario.

Tabla 13.1

Dominios y facetas del Inventario de Personalidad NEO Revisado (NEO PI-R)

NEUROTICISMO	EXTRAVERSIÓN	APERTURA A LA EXPERIENCIA
Ansiedad (N1)	Calidez (E1)	Fantasía (O1)
Enojo hostil (N2)	Gregarismo (E2)	Estética (O2)
Depresión (N3)	Asertividad (E3)	Sentimientos (O3)
Conciencia de sí mismo (N4)	Actividad (E4)	Acciones (O4)
Impulsividad (N5)	Búsqueda de Excitación (E5)	Ideas (O5)
Vulnerabilidad (N6)	Emociones positivas (E6)	Valores (O6)

AGRADABILIDAD	ESCRUPULOSIDAD
Confianza (A1)	Competencia (C1)
Sinceridad (A2)	Orden (C2)
Altruismo (A3)	Responsabilidad (C3)
Condescendencia (A4)	Logro (C4)
Modestia (A5)	Autodisciplina (C5)
Sensibilidad (A6)	Deliberación (C6)

(Adaptado de Costa y McCrae, 1992b, p. 2. Copyright © 1992 por Psychological Assessment Resources, Inc. Reproducido con autorización.)

una forma de informe del observador (Forma R para hombres y Forma R para mujeres) que contiene los mismos 240 reactivos de la forma S presentados en tercera persona. La forma R brinda la posibilidad de obtener valoraciones independientes de los compañeros, cónyuges y otras personas en los mismos dominios que las autoevaluaciones, lo que resulta especialmente importante en el caso del NEO PI-R porque como el inventario supone que el examinado es honesto y está dispuesto a colaborar, no contiene escalas destinadas a verificar la veracidad de las respuestas. En ambas formas se dispone de normas para adultos, hombres y mujeres, y la forma S proporciona normas para hombres y mujeres de edad universitaria.

El modelo de los cinco factores (o de los "Cinco Grandes") postulado por McCrae y Costa ha recibido una aceptación enorme, aunque no universal, como un marco de referencia útil en la exploración de los rasgos de personalidad. Incluso entre los investigadores de la orientación del análisis factorial existen diferencias tanto en el número de los factores de nivel intermedio que proponen como en sus definiciones (L. R. Goldberg, 1993; Zuckerman, Kuhlman, Joireman, Teta y Kraft, 1993). No obstante, en sus diversas versiones el modelo ha estimulado una gran cantidad de investigaciones con el propósito de lograr la identificación cruzada de los factores y la integración de diversos puntos de vista, como los aspectos significativos de las personalidades normal y patológica (véase, por ejemplo, Hofstee, de Raad y Goldberg, 1992). También procede a pasos acelerados el trabajo relacionado en el desarrollo de instrumentos y en el perfeccionamiento de las escalas existentes (Costa y McCrae, 1994, 1995; Costa y Widiger, 1994; Harkness, McNulty y Ben-Porath, 1995; Hogan y Hogan, 1992).

Al evaluar los resultados de esta actividad, debemos tener en cuenta que el análisis factorial sólo proporciona una técnica para reunir los reactivos en grupos relativamen-

te homogéneos e independientes que facilitan la investigación de la validez en comparación con criterios empíricos, permiten una mejor combinación de las puntuaciones para la predicción de criterios específicos y contribuyen a la definición del constructo. La homogeneidad y la pureza factorial son metas deseables en la elaboración de instrumentos, pero no son sustitutos de la validación empírica o de las bases teóricas firmes.

LA TEORÍA DE LA PERSONALIDAD EN EL DESARROLLO DE LAS PRUEBAS

En general, las teorías de la personalidad se originan en medios clínicos, por lo que la verificación experimental a la que han sido sometidas varía de modo considerable de un sistema teórico a otro. Independientemente del grado al que se ha llevado a cabo dicha verificación objetiva, se ha elaborado una serie de tests de personalidad en el marco de una u otra teoría. Como veremos en el capítulo 15, las hipótesis clínicamente formuladas han sido particularmente importantes en la elaboración de las técnicas proyectivas y, aunque en los inventarios autodescriptivos es menos frecuente esta aproximación en la elaboración de instrumentos, disponemos de algunos ejemplos notables.

Inventario Multiaxial Clínico de Millon. Aunque en varios aspectos sigue la tradición del MMPI y en vista de que fue diseñado con los mismos propósitos, el Inventario Multiaxial Clínico de Millon-III (*Millon Clinical Multiaxial Inventory-III*, MCMI-III, Millon, Millon y Davis, 1994), publicado por primera vez en 1977, introduce innovaciones metodológicas importantes. De hecho, se emprendió su preparación con el propósito deliberado de enfrentar las críticas al MMPI y emplear los avances en el diagnóstico de psicopatologías y la elaboración de instrumentos.

El MCMI-III se funda en la postura biopsicosocial de Millon del funcionamiento de la personalidad y la psicopatología (Millon, 1969, 1981, 1990; Millon *et al.*, 1996). Su teoría incluye una matriz de estilos de personalidad derivada de la combinación de tipos en dos dimensiones: la fuente de reforzamiento (es decir, desapegado, discordante, dependiente, independiente y ambivalente) y el patrón de la conducta de afrontamiento (por ejemplo, activa o pasiva). La teoría de Millon de los estilos de personalidad sirvió como una de las bases conceptuales en la formulación original del Eje II de las categorías de trastornos de personalidad del *Manual Estadístico y Diagnóstico de Trastornos Mentales-III* (DSM-III, 1980) preparado por la Asociación Psiquiátrica Estadounidense (*American Psychiatric Association*), ahora en su cuarta edición (DSM-IV, 1994). A su vez, las escalas clínicas del MCMI-III son consistentes, aunque no idénticas, con el sistema de clasificación seguido en el DSM-IV. Los esfuerzos por coordinar el inventario tanto como fuera posible con la evolución del marco de referencia del DSM han fomentado las revisiones frecuentes a que ha sido sometido el MCMI.

El MCMI-III contiene 175 afirmaciones autodescriptivas a las que el examinado debe responder con "cierto" o "falso". El perfil de calificación incluye 24 escalas clínicas, cada una de las cuales se basa en 12 a 24 reactivos traslapados que a menudo aparecen hasta en tres escalas diferentes, aunque con pesos distintos. Los reactivos que cumplen todos los criterios de validación para su escala "original" reciben un peso de 2, mientras que los reactivos secundarios reciben un peso de 1. Como puede verse en la tabla 13.2, las escalas clínicas se agrupan en cuatro categorías principales: Patrones

Clínicos de Personalidad, Personalidad Patológica Grave, Síndromes Clínicos y Síndromes Graves. Las dos primeras categorías contienen escalas diseñadas para evaluar los patrones de trastornos permanentes de personalidad del Eje II del DSM con niveles distintos de gravedad. Las otras dos comprenden algunos de los síndromes del Eje I del DSM. También contiene tres índices de modificación y una verificación de validez diseñada para detectar patrones atípicos de respuesta y sesgos del examinado. Originalmente, el MCMI sólo podía ser calificado por computadora, en la actualidad, además de los servicios de calificación por correo electrónico y de los programas de cómputo para generar los perfiles e informes interpretativos, también es posible la calificación manual, aunque ésta es laboriosa por la necesidad de incorporar transformaciones de las puntuaciones y diversos ajustes.

Una de las innovaciones más significativas del MCMI es el uso de calificaciones estándares denominadas Tasa Base (TB) que, en lugar de ser normalizadas, se fijan en las tasas frecuentes de las características medidas. Las calificaciones TB de corte de las escalas MCMI se establecen de modo que reflejen los datos actuales de la tasa base de poblaciones psiquiátricas para las condiciones particulares que evalúa la escala y mejorar de ese modo el diagnóstico diferencial. Como las tasas dominantes pueden variar en medios y poblaciones clínicas diferentes, las calificaciones TB de algunas de las escalas del MCMI-III pueden ajustarse a ambientes, cronicidad y calificaciones en ansiedad y depresión, así como a ciertos patrones de respuesta.

En el desarrollo de reactivos para el MCMI se siguió la aproximación multifacética que caracteriza a la práctica reciente en la elaboración y validación de los inventarios de personalidad. A este respecto, el MCMI rebasa la metodología descrita en varias

Tabla 13.2

Escalas del Inventario Multiaxial Clínico de Millon-III

Patrones de personalidad clínica
- Esquizoide
- Evasiva
- Depresiva
- Dependiente
- Histriónica
- Narcisista
- Antisocial
- Agresiva (sádica)
- Compulsiva
- Pasiva-Agresiva (negativista)
- Autodestructiva

Patología grave de personalidad
- Esquizotípica
- Fronteriza
- Paranoide

Síndromes clínicos
- Ansiedad
- Somatoforme
- Bipolar: maniaco
- Distimia
- Dependencia del alcohol
- Dependencia de las drogas
- Trastornos de estrés postraumático

Síndromes graves
- Trastorno del pensamiento
- Depresión mayor
- Trastorno alucinatorio

Índices de modificación
- Revelación
- Deseabilidad
- Degradación

Índice de validez

(Adaptado de Millon *et al.*, 1994, p. 2. Copyright © 1994 por DICANDRIEN, INC. Reproducido con autorización.)

secciones de este capítulo. El procedimiento incluye una secuencia de tres pasos principales: (1) teórica-sustantiva (por ejemplo, escribir y seleccionar reactivos adecuados a los constructos clínicamente relevantes), (2) interna-estructural (por ejemplo, correlaciones reactivo-escala y las frecuencias de apoyo) y (3) criterio externo (por ejemplo, diferenciación de los grupos diagnósticos de los de referencia y validación cruzada de nuevas muestras).

Los grupos de referencia empleados para el análisis de reactivos del MCMI-III y sus predecesores no incluían muestras normales sino que estaban formados por pacientes psiquiátricos no diferenciados. Es precisamente por el uso exclusivo de muestras clínicas en la derivación de las normas y puntuaciones transformadas que el autor declara explícitamente que "El MCMI-III no es un instrumento general de personalidad que pueda usarse con poblaciones normales o para cualquier propósito diferente a la detección diagnóstica o a la evaluación clínica" (Millon *et al.*, 1994, capítulo 5). Además, las puntuaciones de corte y la interpretación de perfiles de la escala se ajustan más a quienes muestran psicopatología en los grados medios de gravedad que a aquellos cuyos problemas se acercan a la normalidad o a un grado extremo de gravedad. En algunos de los estudios del MCMI en los que se utiliza a sujetos normales se ha descubierto que éstos califican más alto que los pacientes psiquiátricos, pero a niveles subclínicos, mientras que otros muestran calificaciones normales a niveles patológicos en ciertas escalas (por ejemplo, Histriónica y Narcisista). Tales descubrimientos confirman lo inadecuado de aplicar el inventario a personas normales e indica que, con elevaciones de nivel medio, las escalas pueden, incluso, medir atributos de una personalidad saludable (Wetzler, 1990).

Uno de los propósitos principales del MCMI es contribuir al proceso del diagnóstico diferencial, una tarea que se complica por el hecho de que en el mismo individuo pueden coexistir diferentes condiciones psiquiátricas. Las revisiones y perfeccionamientos relativamente frecuentes por los que ha pasado el MCMI desde su primera publicación resultan loables, pero también hacen más difícil la tarea de evaluar el grado al que ha cumplido sus objetivos; sin embargo, se han acumulado numerosas investigaciones sobre las versiones anteriores del instrumento y empiezan a aparecer algunas publicaciones relacionadas que sin duda serán de utilidad para los usuarios. La investigación ha hecho descubrimientos que indican que el traslape de reactivos entre escalas puede disminuir su poder discriminativo, en especial cuando los examinados se encuentran inmersos en estados depresivos y de ansiedad; aunque se necesitan más datos acerca de la eficiencia diagnóstica de las escalas de síndromes clínicos, el MCMI es un instrumento potencialmente valioso para el diagnóstico de trastornos de personalidad y para la evaluación de su tratamiento (Choca, Shanley y Van Denburg, 1992; Craig, 1993; Goncalves, Woodward y Millon, 1994; Retzlaff, 1995; C. R. Reynolds, 1992a).[10]

Recientemente, Millon preparó dos nuevos instrumentos que amplían sus planteamientos a la evaluación de la personalidad y la psicopatología. Uno es el Inventario Clínico para Adolescentes de Millon (*Millon Adolescent Clinical Inventory*, MACI, Millon, Millon y Davis, 1993), que pretende ser la herramienta a elegir para evaluar a adolescentes de entre 13 y 19 años en medios clínicos. El MACI se formuló a partir del

[10] El Inventario de Personalidad Inadaptada y Adaptada (SNAP) es una nueva medida interesante que también está diseñada para evaluar la personalidad patológica. En contraste con el MCMI, el SNAP emplea escalas derivadas mediante análisis factoriales (véase, por ejemplo, Clark, McEwen, Collard y Hickok, 1993).

Inventario de Personalidad para Adolescentes de Millon (*Millon Adolescent Personality Inventory*, MAPI, Millon, Green y Meagher, 1982), un instrumento anterior destinado originalmente para emplearse lo mismo en la evaluación clínica que en la consejería vocacional y académica, con escalas que evalúan los estilos básicos de personalidad, las preocupaciones expresadas y las tendencias conductuales de los adolescentes.[11] Por otro lado, el Índice de Estilos de Personalidad de Millon (*Millon Index of Personality Styles*, MIPS, Millon, 1994) pretende ser una medida de personalidad para adultos normales que buscan ayuda para resolver problemas laborales, familiares o sociales en diversos ambientes de consejería. El MIPS se estandarizó con muestras de adultos y universitarios y combina elementos de la teoría de personalidad de Millon con las de Freud y Jung.

Inventario de Preferencias Personales de Edwards. Entre las teorías de personalidad que han estimulado el desarrollo de instrumentos, una de las más prolíficas ha sido el sistema de necesidades manifiestas propuesto por Murray y sus colaboradores en la Clínica Psicológica de Harvard (Murray *et al.*, 1938). Uno de los primeros inventarios diseñados para evaluar la fuerza de esas necesidades fue el Inventario de Preferencias Personales de Edwards (*Edwards Personal Preference Schedule*, EPPS, Edwards, 1959). A partir de 15 necesidades derivadas de la lista de Murray, Edwards preparó conjuntos de reactivos cuyo contenido parecía ajustarse a cada una de esas necesidades, entre las que se incluyen el Logro (hacer lo mejor de uno mismo y realizar algo difícil), Deferencia (conformarse con lo que se espera de uno), Exhibición (ser el centro de atención), Intracepción (analizar los motivos y sentimientos propios y ajenos), Dominancia (influir en los demás y ser considerado como líder) y Tendencia a Consolar (ayudar a los demás cuando están en problemas).

El inventario consta de 210 pares de afirmaciones en las que los reactivos de cada una de las 15 escalas se emparejan con los reactivos de las otras 14.[12] En cada par, los examinados eligen una afirmación, la que mejor los caracterice. Es importante considerar que debido a la elección forzada, el EPPS produce puntuaciones *ipsativas*, es decir, la fuerza de cada necesidad no se expresa en términos absolutos, sino en relación con la fuerza de las otras necesidades del individuo, por lo que el marco de referencia de la puntuación ipsativa es del individuo más que la muestra normativa. Como la suma de todas las puntuaciones de la subescala es una constante para todos los examinados, si la puntuación que obtiene el sujeto en una subescala se mueve un punto hacia arriba, la puntuación de otra subescala debe descender un punto. En tales condiciones, dos individuos con puntuaciones idénticas en el EPPS pueden diferir marcadamente en la fuerza absoluta de sus necesidades. Aunque el EPPS proporciona normas para la conversión de sus puntuaciones en percentiles, la naturaleza ipsativa de sus puntuaciones puede cuestionar la pertinencia de dicha conversión. Si bien el marco de referencia ipsativo puede ser el más adecuado para las comparaciones *intraindividuales*, como las que se requieren en la evaluación de los intereses y otras preferencias, se requieren datos normativos de referencia para las comparaciones *interindividuales* que se emplean, por ejemplo, en la evaluación de la anormalidad (Fedorak y Coles,

[11] El MAPI se recomienda ahora como un instrumento que puede usarse únicamente en evaluaciones no clínicas de la personalidad de adolescentes normales.

[12] En la siguiente sección analizaremos esta forma de reactivo, que representa una característica importante del EPPS, como un ejemplo de la técnica de elección forzada.

1979); no obstante, la combinación de ambos marcos de referencia hace que la interpretación de las puntuaciones resulte confusa y menos significativa de lo que podría ser con una aproximación constantemente ipsativa o bien normativa.

Los datos de validez presentados en el manual del EPPS son magros, y aunque se ha publicado un gran número de estudios de validación independientes, a menudo es difícil interpretar sus resultados porque casi ninguno ha tenido en consideración la naturaleza ipsativa de las puntuaciones. Con puntuaciones ipsativas, la media de las intercorrelaciones de las escalas independientes tiende a ser negativa, y la correlación promedio de todas las escalas con cualquier variable externa se aproxima a cero (Hicks, 1970). Estas limitaciones artificiales impiden el análisis apropiado de las puntuaciones ipsativas con los procedimientos correlacionales comunes, por lo que no es de sorprender que los estudios publicados de validación hayan arrojado resultados conflictivos y poco concluyentes (véase, por ejemplo, McCrae y Costa, 1992). A pesar de su simplicidad y de sus rasgos notables, es necesario revisar al EPPS para eliminar las debilidades técnicas relacionadas con la forma de los reactivos así como la interpretación de las puntuaciones.

Forma de Investigación de la Personalidad y otros inventarios de Jackson. La Forma de Investigación de la Personalidad (*Personality Research Form*, PRF) refleja muchos de los avances técnicos en la elaboración de pruebas, que incluyen algunos procedimientos de selección de reactivos que resultaban prácticamente imposibles antes de contar con computadoras de alta velocidad. La PRF ejemplifica el planteamiento de Douglas N. Jackson al desarrollo de tests de personalidad, que comienza con descripciones detalladas y explícitas de los constructos que pretende medir, que constituyen la base para escribir los reactivos y definir los rasgos que los jueces tienen que valorar en los estudios de validación (Jackson, 1970, 1989b).

Se dispone de cinco opciones diferentes de la PRF, que incluyen dos conjuntos de formas paralelas (A,B y AA,BB) con 300 y 440 reactivos, respectivamente. Las formas más largas ofrecen puntuaciones para 22 escalas, cada una de las cuales se basa en 20 reactivos, incluyendo dos puntuaciones de validez, Infrecuencia y Deseabilidad; las formas más cortas contienen sólo 15 escalas de 20 reactivos. Una versión adicional (la E), que se desarrolló más tarde mediante el uso de técnicas más elaboradas para el análisis de reactivos, está formada por 352 de los mejores reactivos de las formas más largas y proporciona puntuaciones en las 22 escalas, aunque cada una consta sólo de 16 reactivos. Además, el vocabulario de la forma E, que es actualmente la de mayor uso, es más sencillo que el de las otras formas. La puntuación de Infrecuencia, diseñada como índice del descuido, los errores en la comprensión de las instrucciones y otras formas no propositivas de respuesta, se basa en el número de respuestas muy improbables que elige el examinado, digamos "Trato de dormir al menos un poco cada noche" y "Yo hago mi propia ropa y zapatos". También se presenta una escala de Deseabilidad, aunque este sesgo ya se había reducido considerablemente gracias a los procedimientos empleados en el desarrollo y la selección de reactivos. El manual advierte correctamente que las puntuaciones demasiado elevadas o bajas en esta escala pueden indicar no sólo actitudes atípicas del examinado (por ejemplo, un intento deliberado por crear una impresión favorable frente a la simulación), sino también características de personalidad que son importantes por derecho propio (digamos, una autoestima elevada o un alto grado de socialización convencional frente a una baja autoestima).

Igual que muchos otros instrumentos de personalidad, la PFR tomó como punto de partida la teoría de la personalidad de Murray; Jackson utilizó la amplia bibliografía teórica y de investigación acumulada a lo largo de tres décadas para formular definiciones conductualmente orientadas y mutuamente excluyentes de 20 constructos o rasgos de personalidad. De ésos, 12 tienen los mismos nombres que los que aparecen en el EPPS, y, para cada una de las 20 escalas bipolares de personalidad, el manual ofrece una descripción de quienes obtienen las puntuaciones más altas y un conjunto de adjetivos que definen los rasgos. En la tabla 13.3, se muestran dos ejemplos de las definiciones de las escalas.

Mediante procedimientos cuidadosamente controlados se generaron bancos de más de 100 reactivos para cada escala y para cada una se eligieron 20 reactivos de acuerdo con una correlación biserial alta con la puntuación de la escala total y una correlación baja con las puntuaciones en otras escalas de rasgos y con la escala de Deseabilidad. Se eliminaron los reactivos cuya proporción flexible era excesiva; mediante un programa especial se asignaron los reactivos a las formas paralelas en términos de la correlación biserial con su propia escala y con la frecuencia de aprobación. Los procedimientos utilizados en la elaboración de la forma E incluyeron, para cada reactivo, el cálculo de un índice de eficiencia de acuerdo con los pesos derivados de varios parámetros estadísticos, lo que permite ordenar los reactivos de esta escala sobre la base de su eficacia.

La validez de constructo de la PRF depende en gran medida de los procedimientos seguidos en el desarrollo y la selección de reactivos para cada escala. Los análisis

Tabla 13.3

Ejemplos de las definiciones de escala para la Forma de Investigación de la Personalidad

Escala	Descripción del examinado que obtiene la puntuación más alta	Adjetivos que definen el rasgo
Estructura cognoscitiva	No le gusta la ambigüedad ni la incertidumbre en la información; desea que todas las preguntas se le respondan completamente; desea tomar decisiones basándose en el conocimiento definido más que en conjeturas o probabilidades	Preciso, exacto, definido, busca la certidumbre, meticuloso, perfeccionista, esclarecedor, explícito, exacto, riguroso, literal, evita la ambigüedad, rígido, determinante, definido, necesita estructura
Percepción	Percibe los olores, sonidos, visiones, gustos y la forma en que se sienten las cosas; recuerda esas sensaciones y cree que son parte importante de la vida; es sensible a muchas formas de experiencia; puede mantener una visión esencialmente hedonista o estética de la vida	Estético, disfruta las sensaciones físicas, observador, mundano, consciente, nota el ambiente, sensible, sensual, abierto a la experiencia, perceptivo, atento, con capacidad discriminativa, atento a las impresiones

(Tomado de Jackson, 1989b, pp. 6-7. Copyright © 1989 por Douglas N. Jackson. Reproducido con autorización.)

factoriales posteriores corroboraron el agrupamiento de reactivos en las 20 escalas. Las correlaciones con escalas comparables del Inventario Psicológico de California, el Estudio de Guilford y Zimmerman sobre el Temperamento, y el NEO-PI, entre otros, ofrecen apoyo adicional a la identificación de los rasgos. Aunque los estudios de las correlaciones entre varios instrumentos diseñados para medir las necesidades de Murray —como la PRF, el EPPS, el TAT (capítulo 15), y la ACL (capítulo 16)— han producido resultados inconsistentes, parece haber más apoyo para la validez de constructo de la PRF que para algunos otros instrumentos, lo que no debería resultar sorprendente al considerar la cuidadosa atención con que se formularon los rasgos en el desarrollo de dicho instrumento (Costa y McCrae, 1988; D. W. Fiske, 1973; Rezmovic y Rezmovic, 1980). Los datos sobre la validez empírica de la PRF comparada con las valoraciones de los compañeros y las autodescripciones también han arrojado resultados prometedores. Un hecho notable es que se ha demostrado que la PRF puede aplicarse a culturas diferentes, incluso a las no occidentales (Jackson, Guthrie, Astilla y Elwood, 1983; Paunonen, Jackson, Trzebinski y Forsterling, 1992). En conjunto, como atestigua la extensa bibliografía compilada por MacLennan (1992), la PRF es un excelente instrumento de investigación, pero aún se necesita mayor información para determinar su eficacia en situaciones prácticas.[13]

El Inventario de Personalidad de Jackson, Revisado (*Jackson Personality Inventory*, JPI-R) se preparó después de la PRF con procedimientos de elaboración de escalas similares aunque perfeccionadas y con una orientación más práctica (Jackson, 1976, 1994a). Las escalas de rasgos se tomaron de la bibliografía de la psicología social y de la personalidad por su relevancia para la predicción del comportamiento de sujetos normales en diversos contextos. Entre los rasgos cubiertos por las 15 escalas se encuentran ansiedad, cooperación, responsabilidad, astucia social y tolerancia. Los datos sobre su validez se obtuvieron no sólo mediante correlaciones con las valoraciones de los compañeros y autodescripciones, usando el modelo de la matriz de rasgos y métodos múltiples, sino también mediante estudios de grupos particulares de los cuales se disponía de datos conductuales relevantes en contextos reales. En la reciente revisión de este inventario, y como forma de apoyar su uso lo mismo en la consejería académica que en medios laborales, se actualizaron las normas para los estudiantes universitarios y se establecieron nuevas normas para obreros y oficinistas. Además, se realizaron algunas modificaciones técnicas y cambios menores en las escalas que ahora se reúnen en cinco agrupamientos de orden superior que son compatibles con las categorías del Modelo de los Cinco Factores (MCF). Específicamente, las dimensiones Extravertido, Confiable y Analítico del JPI-R son muy similares a las dimensiones de Extraversión, Escrupulosidad y Apertura a la Experiencia del MCF, respectivamente. Además, la dimensión Emocional del JPI-R parece representar una combinación de Neuroticismo y Agradabilidad del MCF. Sin embargo, la dimensión de Oportunista del JPI-R, que se derivó de las escalas de Astucia Social y de Aceptación de Riesgos, no tiene un vínculo directo con el MCF y, por consecuencia, se considera como un factor único.

Volviendo a la evaluación de la psicopatología, en la elaboración del Inventario Básico de Personalidad (*Basic Personality Inventory*, BPI, Jackson, 1989a) Jackson empleó

[13] Este instrumento ha sido revisado ampliamente en los Anuarios de Medición Mental: 7ª MMY, No. 123; 8ª MMY No. 643; 10ª MMY, No. 282; véase también TIP-IV para una lista de referencias actualizada.

los mismos criterios que aplicó al preparar el PRF y el JPI. El BPI, desarrollado a lo largo de 15 años, pretende recrear la eficiencia diagnóstica asociada con el MMPI mediante el uso de escalas que son más sustanciales en términos de contenido, pureza psicométrica y margen de aplicabilidad. Aunque el BPI requiere de normas más representativas, en especial para adultos, ya ha mostrado ser prometedor para el uso clínico en el área de la delincuencia juvenil (Holden y Jackson, 1992; para una revisión, véase Urbina, 1995).

Es probable que en esta revisión de los inventarios autodescriptivos el lector haya advertido una tendencia a la combinación de diversas aproximaciones. Tal es el caso de los inventarios de Jackson y de otros instrumentos desarrollados en la última década, ya que en la selección de los reactivos utilizan todas las estrategias excepto la clave empírica de criterio. Además, hasta los inventarios de Minnesota, que son el ejemplo más destacado del uso de una clave empírica de criterio, contienen ahora escalas que se basan en el análisis de contenido y factorial. Hay evidencias que indican que las escalas de los inventarios de personalidad formuladas mediante alguno de los cuatro métodos principales descritos en este capítulo pueden ser todas eficaces, al menos en lo que concierne a la validez convergente y predictiva (Burisch, 1986); sin embargo, los procedimientos basados en el contenido y en la teoría son más fáciles de elaborar, de mejor uso y con mayor probabilidad de mostrar validez discriminativa que el método de clave empírica de criterio, que a su vez es menos compatible con el resto. Además, existe un amplio acuerdo en el sentido de que: (a) el desarrollo de inventarios debería empezar con una definición explícita del rasgo o constructo que pretende medirse, y (b) la estructura organizacional descrita por Campbell y Fiske en su matriz de rasgos y métodos múltiples (véase el capítulo 5) representa una estrategia óptima para el examen de la validez de constructo de los inventarios de personalidad (Angleitner, John y Löhr, 1986; Hogan y Nicholson, 1988; Ozer y Reise, 1994).

LAS ACTITUDES DEL EXAMINADO Y LOS SESGOS DE RESPUESTA

Falseamiento y deseabilidad social. Los inventarios autodescriptivos son especialmente susceptibles a la posibilidad de una descripción falsa. A pesar de los comentarios iniciales que afirman lo contrario, para la mayor parte de sus reactivos hay una respuesta que puede reconocerse como socialmente más deseable o aceptable que las otras. En esas pruebas, quienes las responden pueden sentirse motivados a "falsear para quedar bien", es decir, elegir respuestas que generen una impresión favorable, como al solicitar empleo o al buscar la admisión en una institución educativa. En otras circunstancias, los individuos pueden sentirse motivados a "falsear para quedar mal", tratando de aparentar un trastorno psicológico mayor al que presentan, por ejemplo, en la evaluación de quienes están sometidos a un juicio por haber cometido un crimen.

Las evidencias del éxito con que esas personas pueden falsear en los inventarios de personalidad son copiosas (véase, por ejemplo, Jacobs y Barron, 1968; Radcliffe, 1966; Stricker, 1969; J. S. Wiggins, 1966). Una demostración común en el salón de clases consiste en pedir a grupos diferentes que asuman ciertos papeles; por ejemplo, a una sección del grupo se le da instrucciones para que conteste cada pregunta como lo haría un estudiante universitario feliz y bien adaptado, mientras que a otra sección se le pide que conteste de la manera en que lo haría una persona gravemente trastornada, y a los miembros de la última sección se les pide que contesten en forma honesta y se re-

fieran a su propia conducta. O bien, las mismas personas pueden presentar dos veces la prueba, primero con las instrucciones de simular de una manera especificada y luego en condiciones comunes. Los resultados de los estudios demuestran a las claras la facilidad con que pueden crearse deliberadamente las impresiones deseadas. Resulta interesante advertir que es posible realizar con éxito incluso el falseamiento específico para un objetivo vocacional (Wesman, 1952). La investigación actual indica que, en particular en los inventarios autodescriptivos, la "validez de facie"[14] de un reactivo incrementa su susceptibilidad al falseamiento, tanto en condiciones realistas como en condiciones controladas de laboratorio. Entre mayor sea la facilidad con la que los individuos pueden identificar (a partir de la redacción del reactivo) el rasgo que pretende evaluarse, más a menudo darán la respuesta deseable (Bornstein *et al.*, 1994).

La tendencia a elegir las respuestas socialmente deseables en los inventarios autodescriptivos no indica por fuerza un engaño deliberado por parte de quienes responden. A. L. Edwards (1957), el primero en investigar la variable de la deseabilidad social, la explicó principalmente como un efecto de apariencia, o la tendencia a "falsear para causar una buena impresión", de la que el individuo puede no ser consciente. Esta tendencia puede indicar una falta de percepción de las propias características, autoengaño o poca disposición a enfrentar las propias limitaciones. Otros investigadores (Crowne y Marlowe, 1964; N. Frederiksen, 1965) presentaron evidencias que indican que la fuerza de la respuesta de deseabilidad social se relaciona con la necesidad más general que tenga el individuo de protegerse, de evitar las críticas, de conformidad y de aprobación social. Por otro lado, el individuo que en una autodescripción elige respuestas poco favorables puede estar motivado por la necesidad de recibir atención, simpatía o ayuda para enfrentar los problemas personales; por ejemplo, en los inventarios de personalidad, quienes buscan psicoterapia tienen más probabilidad de parecer más desadaptados de lo que es realmente el caso.

Asimismo, no puede suponerse que la investigación básica esté libre de los efectos de las disposiciones de respuesta; por ejemplo, en el contexto de los estudios sobre el cambio de actitud, los investigadores han demostrado que los resultados pueden verse afectados por condiciones como la percepción que tengan los participantes de lo que espera el experimentador, su afán por complacerlo o frustrarlo y el deseo de proteger su propia imagen (Silverman y Shulman, 1970). Algunas diferencias no sospechadas en estas disposiciones de respuesta pueden explicar en parte la imposibilidad de reiterar los resultados al repetir los experimentos.

Algunos investigadores (Paulhus, 1984, 1986; Paulhus y Reid, 1991) han subrayado la diferencia entre las nociones de manejo de impresiones[15] y autoengaño como explicaciones de las respuestas socialmente deseables o indeseables. El *manejo de impresiones* se refiere a la simulación consciente que se hace con el propósito de crear algún efecto específico, deseado por el que responde; puede considerarse como un contaminante de los datos autodescriptivos y como algo que tiene que ser evaluado y, cuando sea posible, minimizado o evitado. Por otro lado, resulta mucho más complejo el autoengaño, que consiste en una forma sesgada de responder, por lo general de manera positiva, que el individuo honestamente considera como cierta. El autoengaño se

[14] Véase el capítulo 5.

[15] En ocasiones, se emplean "manipulación" y "engaño a los demás" en lugar de "manejo de impresiones".

vincula con otros conceptos relacionados con el yo, como la autoimagen y la autoestima, y con la idea psicoanalítica de mecanismos de defensa. Es una variable digna de estudio por derecho propio como posible indicador de un buen ajuste, hasta cierto punto, y predictiva de otros criterios independientes; por ejemplo, se ha advertido que algunas escalas autodescriptivas pueden arrojar resultados que muestran una "salud mental ilusoria" en negadores defensivos que se esfuerzan por mantener la creencia en su propio ajuste (véase, por ejemplo, Shedler, Mayman y Manis, 1993). Por lo tanto, la relación entre el autoengaño y el ajuste no es simple ni directa, y, para mayor complicación, parece que ciertas características lingüísticas de los reactivos interactúan con las variables del individuo para producir respuestas sesgadas (Helfrich, 1986).

Se han empleado diversos procedimientos en el esfuerzo por enfrentar en los inventarios de personalidad el problema del falseamiento y de las predisposiciones de respuesta. La formulación de reactivos relativamente "sutiles" o socialmente neutros puede reducir la operación de esos factores en algunos inventarios; sin embargo, como Jackson (1971) señalaba, la validez de esos reactivos para la dimensión considerada suele ser poca. Si los examinados están convencidos de que les conviene responder honestamente, las instrucciones del instrumento y el establecimiento de *rapport* deberían motivarlos para responder de manera franca, y en muchas situaciones así ocurre; empero, en otras el método es ineficaz y quizá no tenga mucho efecto sobre la predisposición de las respuestas a la deseabilidad social de las que el individuo no está consciente.

Otra aproximación a la evaluación de respuestas socialmente deseables y de otras formas de manejo de impresiones consiste en establecer escalas especiales que puedan incluirse en un inventario o aplicarse por separado en una batería de pruebas. Una de las primeras fue la escala de deseabilidad social de Edwards (1957), que se desarrolló seleccionando reactivos en función del acuerdo de los jueces sobre sus valores muy altos o bajos de deseabilidad social. Otras, como la escala de Buena Impresión del CPI y la escala SD de J. S. Wiggins (1959), seleccionaron los reactivos según las diferencias en las frecuencias de apoyo que recibían de individuos que respondían en condiciones de "falsear para quedar bien" y otros que respondían en condiciones normales. El tercer método lo ejemplifica la escala de Mentiras del MMPI, que contiene reactivos escritos con el propósito de que sean contestados en la dirección socialmente deseable sólo por los examinados que muestran un sesgo positivo irreal. Se han elaborado otras medidas con el propósito específico de detectar el falseamiento y las respuestas descuidadas o al azar, como la escala F del MMPI.[16] El uso de reactivos de elección forzada es otro procedimiento que se orienta no a la detección, sino más bien a la prevención de la simulación.

La técnica de elección forzada. En esencia, la técnica de elección forzada requiere que el examinado elija entre dos términos o frases descriptivas que parecen ser igualmente aceptables pero que difieren en su validez. Las parejas de frases pueden ser o no deseables, aunque los reactivos de elección forzada también pueden contener tres, cuatro o cinco términos, en cuyo caso los individuos deben indicar qué frase es la que más y cuál la que menos los caracteriza. Otra variante requiere que se haga una elección entre dos respuestas contrastantes en el mismo rasgo, que se califican en una sola escala. Aunque se emplea poco en los inventarios de personalidad, esta forma de reac-

[16] Véase Berry, Wetter y Baer (1995) para una revisión cuidadosa de las escalas para la evaluación del falseamiento.

tivo tiene la ventaja de producir puntuaciones normativas más que ipsativas, por lo que no impone limitaciones artificiales sobre las correlaciones de las diferentes escalas. El indicador de Tipos de Myers-Briggs, que analizaremos en el capítulo 16, ofrece un ejemplo de esa técnica.

El uso de la técnica de elección forzada para controlar la deseabilidad social requiere dos tipos de información sobre cada alternativa de respuesta, su deseabilidad social o "índice de preferencia" y su validez o "índice discriminativo". Este último puede determinarse sobre la base de cualquier criterio específico que el inventario pretenda predecir, como el aprovechamiento académico o el éxito en cierto empleo; aunque también puede basarse en la carga factorial de los reactivos o en su relevancia teórica para diferentes rasgos. La deseabilidad social puede encontrarse al hacer que un grupo representativo valore los reactivos para esta variable o al averiguar la frecuencia con que son aprobados en las autodescripciones. Se ha demostrado que existe una correlación muy alta entre la frecuencia de elección y la deseabilidad social juzgada (Edwards, 1957); en otras palabras, la autodescripción *promedio* de una población concuerda con la descripción que en promedio hace de una personalidad deseable.

Aunque los reactivos de elección forzada pueden reducir la influencia de la deseabilidad social, no debe suponerse que la eliminen por completo. Cuando los reactivos del EPPS se presentan en un formato de elección libre, las puntuaciones correlacionan muy alto con las obtenidas en el formato de elección forzada de la misma prueba (Lanyon, 1966). Más aún, la deseabilidad social juzgada de reactivos particulares no es constante para todos los propósitos, sino que puede diferir para distintas ocupaciones. Así, una prueba de elección forzada en la que se igualó la deseabilidad social general de sus reactivos puede incluir falseamiento cuando se aplica a quienes solicitan un empleo, a quienes pretenden ingresar a escuelas profesionales y a otros grupos con una orientación específica. Desde otro ángulo, se ha encontrado que cuando los reactivos se emparejan sobre la base del promedio de los juicios *grupales* sobre la deseabilidad social general pueden estar muy lejos de ser igualados para *individuos* (N. Wiggins, 1966).

En conclusión, parece que la técnica de elección forzada no ha demostrado ser tan eficaz como se esperaba para controlar el falseamiento o la tendencia a la deseabilidad social. Al mismo tiempo, el formato de reactivos de elección forzada, en particular cuando produce puntuaciones ipsativas, genera otras dificultades técnicas y elimina información acerca de la fortaleza absoluta de las características individuales que pueden ser de gran importancia para algunas situaciones de evaluación.

Disposiciones y estilos de respuesta. La tendencia a elegir alternativas de respuesta en función de la deseabilidad social es sólo una de las diversas predisposiciones identificadas en las respuestas a los inventarios autodescriptivos (Lanyon y Goodstein, 1982, pp. 158–169). Aunque la voluminosa bibliografía sobre la operación de las predisposiciones de respuesta en los inventarios de personalidad data de la década de los cincuenta, investigadores anteriores observaron su influencia sobre los test de habilidad y los de personalidad (véase Block, 1965, capítulo 2). Una de las predisposiciones de respuesta que atrajo inicialmente la atención fue la *aquiescencia*, o la tendencia a responder "cierto" o "sí". La aquiescencia se consideraba una variable continua: en un extremo de la escala se encuentran los que de continuo responden afirmativamente y en el otro los que acostumbran responder de manera negativa (Couch y Keniston,

1960). En lo que respecta a la elaboración de inventarios de personalidad, estas predisposiciones implican que el número de reactivos en que una respuesta de "sí" o "cierto" tiene una valoración positiva debe ser igual al número de reactivos en los que la respuesta de "no" o "falso" recibe tal valoración. Este equilibrio se consigue mediante la selección adecuada de los reactivos o su reformulación, como se hizo en la PRF y como ahora se hace con casi todos los nuevos inventarios.[17]

Otra predisposición de respuesta es la *desviación* o tendencia a dar respuestas inusuales o poco comunes. Berg (1967) propuso esta hipótesis y demostró su operación con contenidos no verbales en un instrumento desarrollado especialmente que requiere expresar la preferencia por ciertas figuras geométricas. Las escalas compuestas por reactivos que quizá obtengan respuesta en una dirección por casi todos los examinados, como la escala de Infrecuencia de la PRF de Jackson, pretenden identificar esos patrones desviados de respuesta. Con todo, el propio Jackson señaló que esas escalas generalmente carecen de relevancia conceptual para los criterios externos, por lo que plantean un problema, en especial en contextos como los laborales, en los que se considera importante la pertinencia de los reactivos del cuestionario. Es por ello que al revisar el JPI se eliminó la escala de Infrecuencia (Jackson, 1994a). La tendencia a utilizar opciones extremas en una escala de valoración (por ejemplo, elegir uno o siete en una escala de siete puntos) también se ha identificado como un posible sesgo de respuesta (Paulhus, 1991).

La investigación sobre las predisposiciones de respuesta como la deseabilidad social, la aquiescencia y la desviación ha pasado por diversas etapas. Al principio se consideraban una fuente de varianza de error o irrelevancia que debía eliminarse de las puntuaciones de las pruebas; luego, llegaron a considerarse como indicadores de características de personalidad más amplias y duraderas que por derecho propio eran dignas de medición (Jackson y Messick, 1958, 1962; J. S. Wiggins, 1962). En esta etapa, era común que se les describiera como *estilos de respuesta* y que a su alrededor se acumulara una gran cantidad de datos empíricos, aunque a la postre esos datos fueron cuestionados desde diversas direcciones (Block, 1965; Heilbrun, 1964; Rorer, 1965). Por ejemplo, Block (1965) presentó sólidas evidencias a favor de una interpretación orientada al contenido de los dos principales factores que generalmente se encuentra que pueden explicar la mayor parte de la varianza común de las escalas del MMPI, cuyos exponentes de las predisposiciones y estilos de respuesta se han interpretado como deseabilidad social y aquiescencia.

La controversia sobre las predisposiciones de respuesta y del contenido en comparación con el estilo en la evaluación de la personalidad no ha terminado de resolverse (Edwards, 1990; Hogan y Nicholson, 1988; Jackson y Paunonen, 1980).[18] La mayoría de los investigadores y autores de pruebas concuerda en que es probable que las puntuaciones de los inventarios de personalidad reflejen una combinación de autoengaño, manejo de impresiones y una autoimagen realista, y en que el peso de cada uno de estos componentes ha de variar con el individuo y la ocasión; sin embargo, algunos consideran posible que las tentativas por mejorar la confiabilidad de los datos

[17] Helmes y Reddon (1993) también señalaron que, si una escala es bipolar y su clave de reactivos no está balanceada, disminuye sustancialmente la cantidad de información que transmiten las puntuaciones bajas.

[18] Véase también otros artículos que siguen a Edwards en la misma sección de Comentarios del *American Psychologist*, 1990, pp. 289-295.

de la autodescripción mediante escalas y reactivos especiales resulten contraproducentes, puesto que pueden reducir la validez de las escalas, en especial en el caso de las muestras normales en comparación con las patológicas. Los autores proponen el uso de las habilidades clínicas tanto para garantizar la cooperación del paciente como para interpretar los resultados, así como la inclusión de valoraciones de personas informadas siempre que exista razón para sospechar de una grave distorsión (véase, por ejemplo, Costa y McCrae, 1992a).

El resto de los clínicos, en particular los que participan en la evaluación de psicopatologías, en su mayoría continúa empleando las llamadas escalas de "validez", aunque tiene claro que pueden reflejar estilos y características de personalidad. De hecho, algunos de los instrumentos más recientes y técnicamente más avanzados para la evaluación de la psicopatología, como el BPI de Jackson o el PAI de Morey, utilizan una clave equilibrada de reactivos y escalas especiales para detectar disposiciones de respuesta invalidantes. También existe una nueva variedad de esas escalas, ejemplificadas por las escalas INVAR e INVER del MMPI-2 y del MMPI-A, que hacen uso de pares de reactivos especialmente seleccionados cuyo contenido es similar u opuesto para detectar respuestas inconsistentes o contradictorias. Por la forma en que están constituidas, es poco probable que las escalas INVAR e INVER, que son similares a la escala de Descuido de Greene (1978) para el MMPI original, se confundan con la varianza de un rasgo válido de personalidad (Ozer y Reise, 1994).

En cualquier caso, el argumento sobre las predisposiciones y estilos de respuesta ha estimulado la investigación y producido cientos de publicaciones. Igual que muchas controversias científicas, su efecto neto ha consistido en aumentar nuestra comprensión de los problemas metodológicos y, por ende, mejorar la elaboración de los inventarios de personalidad y su uso en ambientes aplicados y de investigación.

RASGOS, ESTADOS, PERSONAS Y SITUACIONES

Interacción entre personas y situaciones. A finales de los sesenta y durante los setenta alcanzó su punto más álgido una larga controversia sobre la generalización de los rasgos psicológicos contra la especificidad situacional del comportamiento. Algunos adelantos ocurridos en los sesenta concentraron la atención en algunas "conductas de interés" apenas definidas, distanciadas de los rasgos ampliamente definidos con más generalidad. En el área de las habilidades, ilustran el tema los programas de instrucción individualizada y las evaluaciones referidas a dominio (capítulo 3), así como el diagnóstico y tratamiento de los problemas de aprendizaje (capítulo 17). En el campo de la personalidad, el mayor impulso hacia la especificidad conductual en la evaluación provino de las teorías cognoscitiva social y del aprendizaje social, que son el cimiento de la modificación de conducta y la terapia conductual (Bandura, 1969, 1986; Goldfried y Kent, 1972; Mischel, 1968, 1969, 1973). Las críticas se dirigían sobre todo contra la visión inicial de los rasgos como entidades causales subyacentes que eran fijas e inmodificables, crítica que ya se anticipaba en la investigación inicial y en los primeros trabajos de varios psicólogos en relación con todos los rasgos, cognoscitivos y no cognoscitivos (véase el capítulo 11), y aunque algunos psicólogos habían asumido esta concepción extrema de los rasgos, durante el apogeo del argumento de la especificidad situacional era difícil encontrar a alguien que se identificara como "teórico de los rasgos" (Jackson y Paunonen, 1980).

En particular, la especificidad situacional es mucho más característica de los rasgos de personalidad que de las habilidades; por ejemplo, una persona puede ser muy sociable y comunicativa en la oficina, pero tímida y reservada en las reuniones sociales; o un estudiante que hace trampa en los exámenes puede ser escrupulosamente honesto en cuestiones monetarias. Mischel (1968) y D. Peterson (1968) reunieron numerosas evidencias empíricas de que los individuos muestran una considerable especificidad situacional en muchas dimensiones no intelectuales, como agresión, conformidad social, dependencia, rigidez, honestidad y actitudes hacia la autoridad. La mayor consistencia transituacional de las funciones cognoscitivas en comparación con las no cognoscitivas puede explicarse en parte por el hecho de que los antecedentes de las reacciones del individuo muestran mayor estandarización en el dominio intelectual que en el de la personalidad (Anastasi, 1958, capítulo 11; 1970, 1983a); por ejemplo, mientras que el programa escolar contribuye al desarrollo de habilidades cognoscitivas de gran aplicación en las áreas verbal y numérica, el desarrollo de la personalidad ocurre en condiciones mucho menos uniformes. Más aún, en el área de la personalidad la misma respuesta puede conducir a consecuencias sociales que son positivamente reforzantes en una situación y negativamente en otra, por lo que el individuo puede aprender a responder de maneras muy distintas en diferentes contextos. Tales desigualdades en las experiencias, lo mismo en individuos que en situaciones y culturas, puede dar lugar a una mayor ambigüedad en los reactivos presentados en los tests de personalidad que en los que se incluyen en los instrumentos cognoscitivos, de modo que la misma respuesta a determinada pregunta en un inventario de personalidad (que puede formularse como una "situación") puede tener un significado diferente con cada persona.

Debe advertirse que la cuestión de la relación entre rasgos y situaciones también se vincula a la conocida controversia entre herencia y ambiente (D. C. Rowe, 1987). Es probable que las influencias hereditarias se manifiesten mediante rasgos individuales relativamente duraderos que pueden incluir un rasgo de adaptabilidad a las exigencias situacionales. Las influencias ambientales pueden contribuir a la varianza (o la especificidad) situacional así como a la estabilidad del rasgo debido a que el entorno del individuo puede mostrar una mayor constancia a lo largo del tiempo y las situaciones. Nuestra comprensión de las relaciones entre herencia y ambiente y entre rasgos y situaciones mejorará de seguro con el uso de diseños experimentales apropiados que hagan uso de repetidas mediciones longitudinales y transituacionales.

Es indudable que tanto los debates teóricos como la investigación sobre la interacción entre sujeto y situación han enriquecido nuestra comprensión de las muchas condiciones que determinan el comportamiento individual, a la vez que han contribuido al desarrollo de diseños de investigación más elaborados. Al mismo tiempo, se ha generado un consenso entre quienes respaldan diversas posturas respecto a que las explicaciones de rasgos y situaciones pueden coexistir y que, de hecho, la conducta está determinada por la interacción de variables situacionales y de rasgos. Tal acercamiento resultó particularmente claro en una serie de debates cuidadosos y equilibrados sobre el problema publicados entre los años setenta y los ochenta,[19] de los que surgieron varios

[19] Amelang y Borkenau (1986), Bem y Funder (1978), Endler y Magnusson (1976), Epstein (1979, 1980), Epstein y O'Brien (1985), Hogan, DeSoto y Solano (1977), Kenrick y Funder (1988), Mischel (1977, 1979) y Mischel y Peake (1982). Véase también Anastasi (1983b) para una recapitulación.

puntos dignos de interés. La conducta muestra una considerable estabilidad temporal cuando se mide de manera confiable, es decir, al sumar observaciones repetidas hechas de preferencia por varios observadores que están informados respecto al individuo al que se evalúa. Cuando se estudian muestras aleatorias de personas y situaciones, las diferencias individuales contribuyen más que las situacionales a la varianza total de la conducta. La interacción entre personas y situaciones contribuye tanto o un poco más que las diferencias individuales. Para identificar los rasgos de personalidad es necesario que, al evaluar al individuo en diversas situaciones, se utilicen dimensiones públicamente observables y conductas que sean relevantes a las dimensiones en cuestión y se agreguen los resultados (Epstein, 1980; Kenrick y Funder, 1988). A pesar del relativo consenso alcanzado en muchas de estas cuestiones y de lo que ya se ha aprendido, persisten aún muchas interrogantes y han surgido otras en ambos lados del debate respecto a personas y situaciones (véase, por ejemplo, Funder, 1991).

La persona. El grado de especificidad conductual entre situaciones varía con la gente, y los individuos difieren en el grado al que modifican su conducta para enfrentar las exigencias de cada situación. A este respecto, la inconsistencia moderada indica una flexibilidad eficaz adaptativa mientras que la consistencia excesiva sería indicador de una rigidez inadaptada. Más aún, las situaciones particulares en las que la conducta es consistente pueden variar con las personas. La consistencia entre situaciones está influida por la forma en que los individuos las perciben y categorizan lo que a su vez depende de sus metas, motivos y sentimientos así como de su experiencia en situaciones similares.[20]

Las diferencias individuales en la consistencia son de gran interés para los psicólogos, ya que en la medida en que se les pueda evaluar confiablemente es posible emplearlas como variable moderadora en la predicción del comportamiento. Además, desde el punto de vista psicométrico, las diferencias en el individuo y entre individuos en la consistencia se consideran cruciales para moderar la validez y la confiabilidad de toda medición psicológica. Por lo tanto, no resultan sorprendentes los esfuerzos por establecer los medios de evaluar tales diferencias. Uno de éstos se basa en las valoraciones que la gente se asigna en diversos rasgos; en este caso, las puntuaciones se derivan de la varianza de cada persona. La varianza baja entre reactivos de la escala representa la cualidad de *la peculiaridad del rasgo*, que de hecho ha sido asociada a coeficientes elevados de validez (Amelang y Borkenau, 1986; Baumeister y Tice, 1988). Una aproximación diferente la representa la idea de Lanning (1991) de "escalabilidad", definida como el grado al que, según la evaluación de una autodescripción, el sujeto mantiene entre situaciones un ordenamiento normativo de los reactivos conductuales.

Una propuesta interesante que asocia los conceptos de interacción entre persona y situación con la deseabilidad social, en términos del proceso que fundamenta los reactivos a las preguntas de los cuestionarios de personalidad, es la "teoría del umbral" de respuesta postulada por Jackson (véase, por ejemplo, Helmes y Jackson, 1989; Jackson, 1986b). El modelo se basa en la premisa de que los reactivos del inventario representan un microcosmos de conducta en el mundo real y aplica los métodos de la teoría de res-

[20] Esta concepción de la consistencia conductual proviene de la aproximación ideográfica a la evaluación de la personalidad formulada, entre otros, por Allport (1937) y G. A. Kelly (1963).

puesta al ítem o al contenido de las escalas de personalidad (véase capítulo 7). Jackson propone que, así como es posible utilizar las respuestas de muchos individuos a un solo reactivo para derivar una curva característica del reactivo, también es posible usar las respuestas de un mismo sujeto a muchos reactivos para graficar una curva característica del sujeto, que nos permitiría predecir la probabilidad de que un individuo apoye los reactivos tomando como base la relevancia que la dimensión de la deseabilidad social tenga para él, su umbral o predisposición a responder positivamente a los reactivos en términos de la deseabilidad social y la propia deseabilidad de los reactivos.

La situación. Las situaciones también difieren en las restricciones conductuales que imponen, de modo que podemos predecir con un alto nivel de confianza que en una biblioteca los lectores permanecerán en silencio y que los conductores se detendrán ante la luz roja. De igual modo, es probable que, cualquiera que sea la estructura de sus rasgos, las personas naden en la playa y lean en la biblioteca (aunque algunos individuos pueden leer mientras están en la playa o soñar con la playa mientras están en la biblioteca). Una forma de comprender mejor las limitaciones que las situaciones imponen sobre el comportamiento consiste en estudiar las características de diversos ambientes conductuales. El trabajo de Roger Barker sobre la psicología ecológica, recientemente actualizado, ofrece un prometedor conjunto de herramientas para clasificar los medios conductuales y para describir diversas facetas del entorno (Schoggen, 1989).

Las diferencias transculturales pueden considerarse como una instancia especial y más penetrante de variabilidad situacional. Como tal ofrecen una oportunidad única para estudiar la consistencia e inconsistencia en la conducta de los individuos. Con tal fin, las aproximaciones a la evaluación transcultural que estudiamos en los capítulos 9 y 12 en el contexto de la medición de la habilidad pueden usarse también para investigar otros rasgos psicológicos.

Los tests de personalidad muestran grandes diferencias cuando se aplican a culturas distintas, por lo que cualquier explicación de las diferencias culturales y subculturales requiere de un conocimiento específico de las condiciones y circunstancias prevalecientes en cada grupo. Esto ha sido reconocido en la evaluación de la psicopatología en grupos de culturas diferentes en los Estados Unidos (Malgady, Rogler y Constantino, 1987; Paniagua, 1994); por ejemplo, las diferencias grupales en instrumentos como el MMPI pueden reflejar las diferencias en la interpretación de las instrucciones o de reactivos específicos. Las diferencias culturales en la conducta que se considera socialmente deseable puede a su vez influir en las puntuaciones; en algunos grupos las elevadas puntuaciones en las escalas que evalúan la depresión pueden ser resultado de fuertes tradiciones de autodesaprobación y modestia; en otros, las altas puntuaciones pueden indicar la presencia de problemas emocionales genuinos que surgen de prácticas de crianza infantil, conflictos con los papeles sociales, frustraciones de grupos minoritarios y otras diferencias culturales.

El tema de la evaluación de la personalidad en culturas distintas va más allá de la posibilidad de transportar las pruebas; tiene que ver con la posibilidad de trasladar los sistemas conceptuales, como rasgos y jerarquías de rasgos, que se aplican a la conducta (Guthrie, Jackson, Astilla y Elwood, 1983). Además, como se descubrió inicialmente en la evaluación de la habilidad de dominio, cada vez es mayor el reconocimiento de que algunas dimensiones importantes de la personalidad no son universales. En

consecuencia, además de los esfuerzos por traducir y adaptar los instrumentos anglo-americanos tradicionales para poblaciones diferentes, ahora se elaboran algunos tests de personalidad diseñados para evaluar dimensiones específicas en grupos culturales y subculturales (véase, por ejemplo, Dana, 1993; Lonner y Berry, 1986); sin embargo, aunque ciertos reactivos e inventarios tal vez no puedan traducirse a otras culturas, existen razones para creer que el modelo jerárquico de rasgos puede ser de utilidad para integrar los descubrimientos transculturales. Es decir que, en tanto se comience con la medición de las conductas identificadas como significativas en cada cultura con los instrumentos adecuados a cada una, puede ser posible identificar algunos constructos psicológicos de nivel superior que puedan ser universales o al menos muy generalizables a diversas culturas (Anastasi, 1992c; Díaz-Guerrero y Díaz-Loving, 1990).

Rasgos y situaciones. Los inventarios autodescriptivos para evaluar la *ansiedad ante las pruebas* ilustran el hecho de que rasgos y situaciones no constituyen formas incompatibles de categorizar la conducta (I. G. Sarason, 1980). Un ejemplo es el Inventario de Ansiedad (*Test Anxiety Inventory*,[21] TAI) preparado por Spielberger y sus colaboradores (Spielberger *et al.*, 1980). Este instrumento es en esencia la medida de un rasgo que se define en términos de una clase de situaciones centradas en pruebas y exámenes. Quienes obtienen una calificación elevada en el instrumento suelen percibir las situaciones evaluativas como amenazadoras. El inventario comprende 20 afirmaciones que describen las reacciones antes, durante y después de pruebas y exámenes, en las que se pide a los examinados que indiquen cómo se sienten *generalmente* marcando la frecuencia con que experimentan cada reacción (casi nunca, en ocasiones, a menudo o casi siempre). Algunos ejemplos característicos son: "Me paralizo en los exámenes importantes" y "Cuando presento un examen, me siento inquieto y perturbado". El TAI produce una puntuación total sobre la predisposición a sentir ansiedad en situaciones de prueba, así como subpuntuaciones en dos componentes identificados mediante el análisis factorial: preocupación y emocionalidad. En este contexto, la preocupación se define como una "inquietud cognoscitiva por las consecuencias del fracaso" y la emocionalidad como "reacciones del sistema nervioso autónomo que son provocadas por el estrés evaluativo" (Spielberger *et al.*, 1980, p. 1).

El Perfil de Ansiedad Situacional (Oetting y Deffenbacher, 1980)[22] proporciona un mayor grado de especificación situacional. En este instrumento, los examinados valoran sus reacciones sobre reactivos que cubren sentimientos de ansiedad e interferencia de pensamiento. Ambas puntuaciones se obtienen para cada una de seis situaciones de prueba en las que se dan instrucciones a los examinados para que se imaginen en circunstancias que van de "pruebas de opción múltiple" y "pruebas no anunciadas" a "dar una conferencia".

Debido quizá a lo común de la ansiedad ante las pruebas y a la relativa facilidad para realizar investigaciones en los medios académicos, la investigación y teorización sobre el tema continúan con el mismo vigor en los Estados Unidos y en otros países

[21] La forma que se entrega al examinado lleva el título "Test Attitude Inventory" (*Inventario de Actitudes ante las Pruebas*) y así aparece en la novena edición del MMY y en el TIP-IV.

[22] El Perfil de Ansiedad ante las Pruebas ya no está disponible; sin embargo, si el lector está interesado en el instrumento para propósitos de investigación, puede obtener una licencia para reproducirlo con sólo escribir a Eugene Oetting, Department of Psychology, Colorado State University, Fort Collins, CO 80523.

(Hagtvet y Johnsen, 1992). En cualquier caso, el constructo ofrece un buen modelo de la forma en que los rasgos y los conceptos situacionales pueden ser de utilidad para categorizar la conducta, especialmente en el área de la personalidad. Dependiendo del propósito del instrumento, los constructos de rasgos pueden definirse con diferentes grados de amplitud o restricción y asociarse a categorías específicas de situaciones.

Rasgos y estados. Otra forma de conceptuar el dominio de conducta evaluada por los tests de personalidad comprende la diferenciación entre rasgos y estados. El Inventario de Estados y Rasgos de Ansiedad (*State-Trait Anxiety Inventory*, STAI), formulado por Spielberger y sus colaboradores (Spielberger, 1985; Spielberger *et al.*, 1983), ejemplifica mejor esta diferenciación. En la elaboración del instrumento, se definió el estado de ansiedad (E-Ansiedad) como una condición emocional transitoria caracterizada por sentimientos subjetivos de tensión y preocupación. Estos estados varían en intensidad y fluctúan con el tiempo. El estado de ansiedad se mide con 20 afirmaciones descriptivas a las que el individuo responde, haciendo referencia a cómo se siente *en ese momento* (por ejemplo, "Me siento tranquilo", "Estoy nerviosa"). Las respuestas se registran con la indicación de la intensidad del sentimiento (en absoluto, algo, moderadamente, mucho).

El rasgo de ansiedad (R-Ansiedad) se refiere a la predisposición relativamente estable a la ansiedad, es decir, la tendencia del individuo a responder a situaciones que percibe como amenazadoras con una elevada intensidad del estado de ansiedad. Se da instrucciones a los individuos para que indiquen cómo se sienten *generalmente*, marcando la frecuencia en cada una de las 20 afirmaciones (casi nunca, en ocasiones, a menudo, casi siempre). Algunos ejemplos de esas afirmaciones son "me inclino a tomar mal las cosas" y "soy una persona estable". Los individuos con un alto rasgo de ansiedad suelen mostrar elevaciones en el estado de ansiedad más a menudo que las personas en las que el rasgo es bajo porque reaccionan a una gama mayor de situaciones como si fueran amenazadoras o peligrosas; además, son especialmente sensibles a las situaciones personales que suponen cierta amenaza para su autoestima, como la evaluación del desempeño o la experiencia de fracaso; sin embargo, que el estado de ansiedad aumente o no en determinada situación depende del grado al que el individuo perciba la situación como amenazadora o peligrosa sobre la base de su experiencia. El STAI y su versión para menores, el Inventario de Estados y Rasgos de Ansiedad para Niños (*State-Trait Anxiety Inventory for Children*, STAIC) ha sido traducido a 43 idiomas y dialectos y ha acumulado una bibliografía de más de 6 000 estudios (Spielberger, 1989; Spielberger y Sydeman, 1994).

La diferenciación entre estado y rasgo fue aplicada por Spielberger y sus colaboradores en otro inventario desarrollado subsecuentemente, el Inventario de Estados y Rasgos de Expresión de la Ira (*State-Trait Anger-Expression Inventory*, STAXI; Spielberger, 1988; Spielberger, Johnson, Russell, Crane, Jacobs y Worden, 1985). El STAXI comprende 44 reactivos que representan los dominios de experiencia y expresión de la ira. El dominio de la experiencia de la ira es evaluado por dos escalas que son paralelas a las del STAI: el estado de ira (E-Ira) y el rasgo de ira (R-Ira). Este último tiene a su vez dos subescalas, cada una compuesta por cuatro reactivos: temperamento de ira y reacción de ira. La frecuencia de la expresión de la ira es evaluada con tres escalas de ocho reactivos: ira interna, ira externa y control de la ira. En la undécima edición del MMY pueden encontrarse las revisiones que Biskin (1992) y Retzlaff (1992) hicieron de este instrumento.

ESTADO ACTUAL DE LOS INVENTARIOS DE PERSONALIDAD

Además de los problemas comunes que enfrentan todos los instrumentos de evaluación psicológica, el desarrollo y uso de los inventarios de personalidad se ha visto acosado por dificultades especiales. El problema del manejo de impresiones es mucho más agudo en la medición de la personalidad que en la evaluación de la aptitud; además, los cambios temporales son mayores en la conducta medida por los tests de personalidad que la que miden los de habilidad. Esto complica la determinación de la confiabilidad del instrumento, ya que es posible que las fluctuaciones temporales aleatorias se confundan con los cambios conductuales más generales y sistemáticos. Incluso en periodos relativamente cortos, no puede suponerse que las variaciones en las respuestas se restrinjan al instrumento y no caractericen el área de conducta no probada que se considera. Un problema relacionado tiene que ver con el hecho de que la especificidad situacional de las respuestas es mayor en el dominio no cognoscitivo que en el que sí lo es.

La década de los noventa ha presenciado el resurgimiento de la investigación que enfrenta las complejidades de la evaluación de la personalidad y busca soluciones innovadoras a estos añejos problemas. El periodo se caracteriza por importantes avances teóricos y metodológicos.[23] Las primeras críticas a la medición de la personalidad indudablemente tuvieron un efecto saludable y en parte estimularon los subsecuentes desarrollos en esta área de la psicometría; sin embargo, debemos tomar precauciones contra el riesgo de que en el celo por erradicar el pensamiento engañoso también se pierdan conceptos adecuados y útiles; por ejemplo, la propuesta que surge ocasionalmente de que los conceptos de rasgos y de la evaluación diagnóstica de la personalidad deben ser descartados indican una definición innecesariamente restringida de ambos términos. El diagnóstico no supone por fuerza que se pongan etiquetas a las personas, el uso de las categorías psiquiátricas tradicionales o la aplicación del modelo médico de "enfermedad". La evaluación diagnóstica debería emplearse como un auxiliar en la descripción y comprensión del individuo, la identificación de sus problemas y la toma de decisiones apropiada. De modo similar, los rasgos se refieren a las categorías en que la conducta debe clasificarse si queremos tratar con ella, sea en el contexto científico o en cualquier otro. La amplitud óptima de las categorías ha de variar de acuerdo con el propósito de la evaluación. El modelo jerárquico de rasgos que presentamos en el capítulo 11 puede utilizarse igualmente bien en la descripción de la conducta no cognoscitiva. En determinadas circunstancias, serán mejores los rasgos relativamente amplios, pero en otras, será necesario evaluar comportamientos estrechos y específicamente definidos.

[23] El trabajo de Broughton, que aplica el concepto del prototipo y la técnica de escalamiento multidimensional a la evaluación de la personalidad, ejemplifica algunos de los desarrollos más interesantes y de rápida evolución metodológica en el campo (Broughton, 1990; Broughton, Boyes y Mitchell, 1993). Desde un punto de vista teórico, el modelo circunflejo ha proporcionado un marco de referencia integrador para diversas tradiciones del área de la conducta interpersonal (véase, por ejemplo, Hofstee *et al.*, 1992; J. S. Wiggins, 1996; Wiggins y Pincus, 1992). Además, cada vez es mayor el análisis y exploración de las múltiples relaciones entre los dominios de la personalidad y la inteligencia (véase el capítulo 16).

Medición de intereses y actitudes

La naturaleza y la fuerza de los intereses, lo mismo que las actitudes, constituyen un aspecto importante de la personalidad que puede afectar el éxito educativo y ocupacional, las relaciones con los demás, el goce de las actividades recreativas y otras condiciones importantes de la vida cotidiana. Aunque por lo general las pruebas pretenden medir una u otra de esas variables, no es posible hacer una clasificación rígida de los instrumentos en categorías discretas, como los intereses y las actitudes. La superposición es la regla, por lo que puede afirmarse que un cuestionario diseñado para medir la fuerza relativa de los intereses en tareas de investigación, artísticas y ocupacionales también mide las actitudes del individuo hacia la ciencia pura, el arte, las aplicaciones prácticas y cosas similares.

Dada la clara relación de los valores con las elecciones del individuo, a menudo se los estudia junto con los intereses, las actitudes y las preferencias. Aunque los psicólogos sociales y de la personalidad han hecho y continúan realizando muchas investigaciones sobre el tema en las que se incluyen algunos estudios interesantes sobre la generalidad de los valores entre las culturas (S. H. Schwartz, 1992, 1994; S. H. Schwartz y Sagiv, 1995; Super y Šverko, 1995), ha sido magro el avance reciente en el área de los instrumentos estandarizados dirigidos exclusivamente a la medición de los valores. Esto tiene que ver con problemas inherentes a la tarea en cuestión, como las dificultades para hacer un muestreo sistemático al nivel de abstracción apropiado para dominios de valor. Además, algunas de las primeras y más populares pruebas de valores demostraron ser incompatibles con la forma en que, a la postre, se llegó a concebirlos en el campo (Braithwaite y Scott, 1991).[1] Aunque aún se publican o se revisan algunas pruebas "independientes", como es el caso del Inventario de Valores para la Vida (*Life*

Values Inventory, Brown y Crace, 1996) y de la Escala de Valores (*Values Scale*, Nevill y Super, 1989), la evaluación formal de los valores se está incorporando en los instrumentos diseñados para facilitar la toma de decisiones relacionadas con la elección de una carrera y para evaluar las actitudes y los motivos relacionados con el trabajo.[2]

Por su parte, el estudio de los *intereses* recibió su mayor impulso en la consejería educativa y profesional, que veremos en el capítulo 17. Aunque en menor medida, al desarrollo de pruebas en esta área también lo estimuló la clasificación y selección de personal, ya que la consideración de los intereses del individuo reviste importancia práctica desde el punto de vista tanto del empleador como del trabajador. La evaluación de *opiniones* y *actitudes* se originó principalmente como un problema de la psicología social; por ejemplo, son evidentes las implicaciones que tienen para las relaciones grupales las actitudes hacia grupos diferentes. De igual modo, tanto para el psicólogo social como para quienes se dedican a los negocios, la política y otros campos aplicados resultan de profundo interés la medición y predicción de opiniones públicas respecto a una gran variedad de temas, instituciones o prácticas. La medición de opiniones y actitudes también ha progresado en las áreas de investigación del consumidor y de las relaciones con los empleados.

Todos los instrumentos que estudiaremos en este capítulo son inventarios autodescriptivos elaborados mediante alguno de los métodos que explicamos en el capítulo 13; sin embargo, debe advertirse que en esta área, como en la medición de todos los constructos de personalidad, se exploran continuamente otros procedimientos, de los que veremos algunos ejemplos en los capítulos 15 y 16.

INVENTARIOS DE INTERESES: SITUACIÓN ACTUAL

La mayor parte de los inventarios de intereses[3] se diseña para evaluar los intereses individuales en diferentes campos de trabajo. Algunos hacen un análisis de los intereses en los programas educativos o campos de estudio, que a su vez suelen estar vinculados a decisiones relacionadas con la elección de carrera. Aunque la frecuencia de uso de instrumentos en la consejería ha permanecido bastante estable desde la década de los cincuenta, el uso de pruebas de interés se ha incrementado en comparación con el empleo de tests de habilidad (Zytowski y Warman, 1982). Los más recientes inventarios desarrollados o revisados reflejan algunos cambios importantes en la consejería profesional, uno de los cuales tiene que ver con el mayor énfasis en la *autoexploración*. Cada vez son más los instrumentos que permiten que el individuo estudie los resultados detallados de la prueba y los relacione con la información ocupacional y

[1] Por ejemplo, el Estudio de Valores (Allport, Vernon y Lindzey, 1960), uno de los primeros inventarios de valores, consta principalmente de reactivos relacionados con preferencias, creencias e intereses más que con valores; actualmente ya no está en venta.

[2] El lector puede encontrar en Anastasi (1988b, pp. 580–583), Braithwaite y Scott (1991), y en Dawis (1991, pp. 845–850) las descripciones de muchos instrumentos que se relacionan, principal o exclusivamente, con valores.

[3] La Asociación Nacional Estadounidense de Desarrollo de Carrera (*National Career Development Association*, Kapes, Mastie y Whitfield, 1994) publicó una extensa guía, ahora en su tercera edición, sobre los inventarios de intereses y otros instrumentos diseñados para la evaluación profesional. El lector puede encontrar en Borgen (1986) y Hansen (1990) una recopilación de los temas, teóricos y metodológicos, más destacados del campo, así como algunos ejemplos de los mejores instrumentos.

otros datos sobre la experiencia y capacitación personal. También ha aumentado la atención concedida a la preparación para la toma de decisiones relacionadas con la elección de carrera. Los inventarios de intereses que veremos en este capítulo deben considerarse en el contexto de los antecedentes del capítulo 17 sobre la evaluación profesional.

Un segundo cambio sugiere el énfasis que hoy se concede *al aumento de las opciones de carrera* del individuo, y tiene que ver con la medición de los intereses. De hecho, el término "validez de exploración" se utiliza para designar el efecto que los inventarios de intereses pueden tener en términos de aumentar las conductas instrumentales de exploración de las carreras profesionales (véase, Randahl, Hansen y Haverkamp, 1993). Así, los inventarios de intereses, al igual que otros programas más amplios de orientación profesional que veremos en el capítulo 17, se utilizan para familiarizar al individuo con ocupaciones adecuadas que de otra manera no habría considerado.

El tercer cambio significativo también está asociado con la expansión de las opciones de carrera y se relaciona con la imparcialidad sexual o de género de los inventarios de intereses. En general, comparan los intereses que el individuo expresa con los que suelen mostrar las personas que participan en ocupaciones distintas, por lo que se califican las respuestas de determinados reactivos o se interpretan las puntuaciones obtenidas en áreas generales de interés. Aun cuando esta aproximación representa un procedimiento empírico objetivo para evaluar los intereses del individuo, tiende a perpetuar las diferencias grupales en la proporción de hombres y mujeres en algunas ocupaciones, como la ingeniería o la enfermería, diferencias que de una u otra forma suelen influir en la interpretación de los resultados obtenidos por hombres y mujeres en los inventarios de intereses. Por ello, se ha dedicado un gran esfuerzo a la investigación y al debate sobre las formas de reducir los posibles sesgos sexuales en los inventarios de intereses (Tittle y Zytowski, 1978; Zytowski y Borgen, 1983). También se ha preparado y distribuido[4] un conjunto de directrices para la evaluación de sesgos sexuales y la imparcialidad de género en los inventarios profesionales cuya influencia resulta evidente en casi todo inventario reciente o revisado. Uno de los pasos emprendidos ha sido la eliminación de tendencias sexuales en la redacción de los reactivos del inventario; otras soluciones son el equilibrio del contenido de los reactivos en términos de las diferencias en la socialización de roles sexuales y proporcionar las normas más apropiadas posibles para cada grupo sexual en las escalas del inventario, lo que ha contribuido a mejorar la calidad de los datos técnicos de que disponen los consejeros profesionales y otros usuarios de los inventarios de intereses. Sin embargo, debido a la persistencia de las diferencias en los patrones de intereses y en la pertenencia a grupos ocupacionales, es necesario considerar cuidadosamente los datos de los inventarios en la evaluación profesional de las mujeres (Hackett y Lonborg, 1994).

Hace poco surgió en el campo otro desarrollo directamente vinculado a la legislación estadounidense que obliga a proporcionar orientación vocacional a las poblacio-

[4] Preparado como parte del estudio conducido por el Instituto Nacional de Educación (*National Institute of Education*). En Tittle y Zytowski (1978, pp. 151–153) se reproduce una copia.

nes discapacitadas y con desventajas[5] (L. S. Gottfredson, 1986b; Reed, Rotatori y Day, 1990; Szymula, 1990). Entre otras cosas, los requisitos especiales de las personas con discapacidades en términos de la evaluación vocacional han instigado el desarrollo de diversos inventarios pictóricos de intereses que usan dibujos o estímulos audiovisuales, como diapositivas o videocasetes (Elksnin y Elksnin, 1993; Kapes *et al.*, 1994, pp. 307–345). La primera generación de instrumentos diseñados para evitar los requisitos de lectura de los inventarios tradicionales de intereses de lápiz y papel es inadecuada desde el punto de vista psicométrico, en particular en el campo de la validez. Sin embargo, los inventarios pictóricos de intereses representan una forma innovadora de aplicar los estímulos de la prueba que probablemente avance conforme mejore la tecnología y aumente la necesidad de contar con pruebas alternativas para poblaciones especiales. Por supuesto, en la evaluación de los intereses también pueden emplearse las modificaciones de los instrumentos tradicionales, como los que vimos en el capítulo 9 al tratar los tests de habilidad para poblaciones especiales. El impulso proporcionado por la Ley para los Estadounidenses con Discapacidades de 1990 (P.L. 101–336) seguramente estimulará el interés y el trabajo en esta área (Bruyère y O'Keeffe, 1994).

EL INVENTARIO DE INTERESES DE STRONG

Orígenes y desarrollo del Inventario de Intereses de Strong (*Strong Interest Inventory*™, SII). Este inventario, cuya reciente edición es de 1994, tiene una larga historia. El planteamiento general seguido en su construcción fue formulado inicialmente por E. K. Strong, Jr. mientras asistía a un seminario de graduados de 1919 a 1920 sobre la medición de intereses en el Instituto Tecnológico de Carnegie (*Carnegie Institute of Technology*) (D. P. Campbell, 1971, capítulo 11; Fryer, 1931, capítulo 3). Cuando el Formulario de Intereses Vocacionales de Strong (*Strong Vocational Interest Blank*®, SVIB) se publicó por primera vez en 1927, introdujo dos procedimientos principales en la medición de los intereses ocupacionales. Primero, los reactivos abordan el gusto o disgusto de los individuos por una amplia variedad de actividades, objetos o tipos específicos de personas que se encuentran en la vida cotidiana. Segundo, las respuestas se pusieron en clave empírica para diferentes ocupaciones, por lo que estos inventarios estuvieron entre los primeros instrumentos en utilizar la clave de los reactivos como criterio, que más tarde fue seguida en la preparación de inventarios de personalidad como el MMPI y el CPI (capítulo 13). Se descubrió que quienes realizan diferentes ocupaciones se caracterizan por compartir intereses que los distinguen de los que participan en otras actividades. Estas diferencias de intereses no sólo se extienden a los asuntos relacionados directamente con las actividades laborales, sino también a las materias escolares, los pasatiempos, las clases de deportes, los juegos o los libros favoritos, las relaciones sociales y muchas otras facetas de la vida cotidiana. Se demostró así que era posible preparar un inventario que explorara los intereses del individuo por las cosas familiares y determinar qué tan cercano es el parecido de sus intereses con los de quienes se desempeñan con éxito en ciertas ocupaciones.

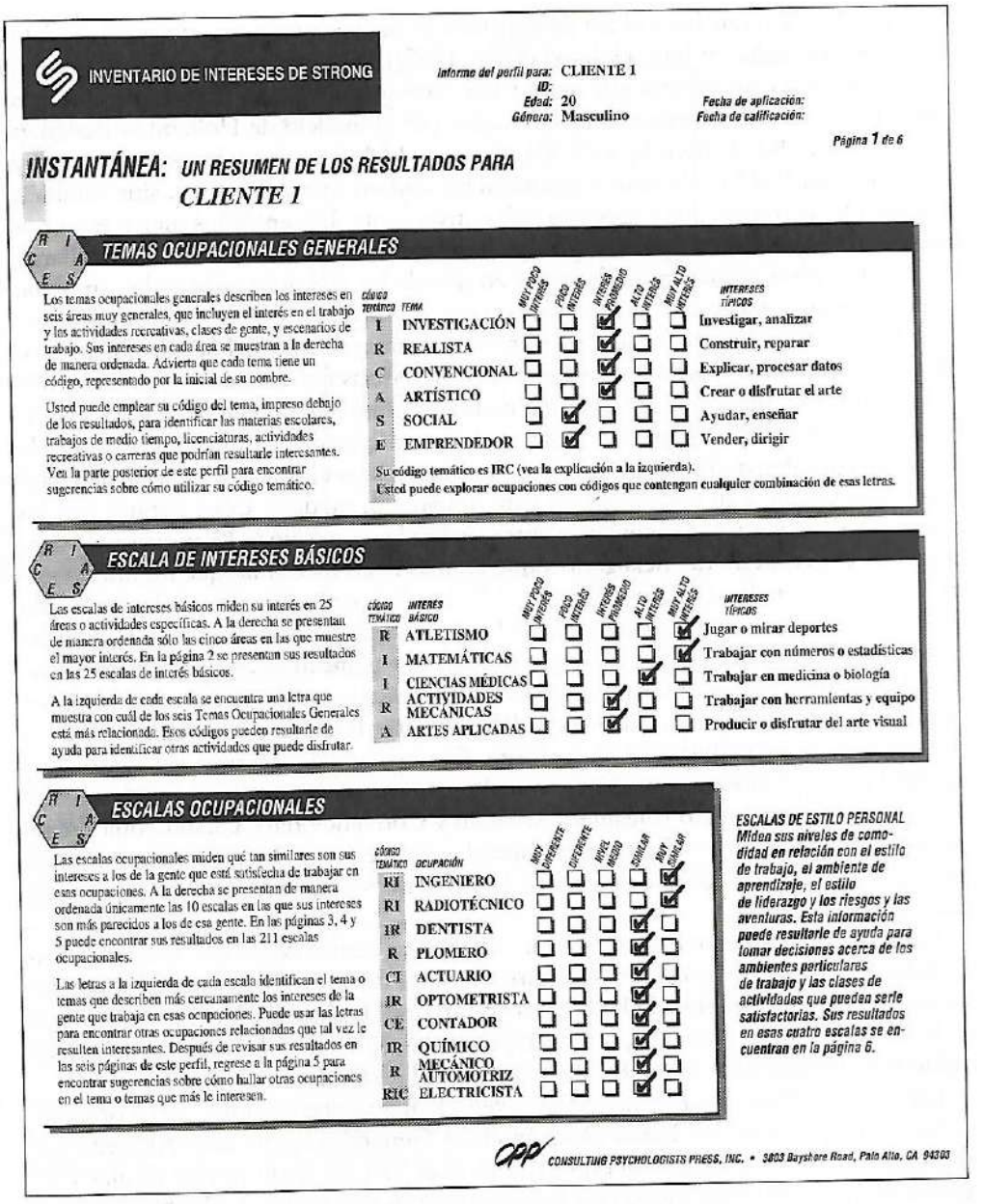
[5] Dos ejemplos específicos de esa legislación son la Ley de Educación Vocacional Carl D. Perkins, de 1984 (P.L. 98–524) y Las Enmiendas de 1990 a la Ley de Educación Vocacional y Tecnología Aplicada Carl D. Perkins (P.L. 101–392).

Desde la década de los setenta han tenido lugar grandes innovaciones en las revisiones sucesivas del SII (D. P. Campbell, 1974; D. P. Campbell y Hansen, 1981; Hansen y D. P. Campbell, 1985; Harmon, Hansen, Borgen y Hammer, 1994). Los principales cambios comprenden: (1) la introducción de un marco teórico para guiar la organización e interpretación de las puntuaciones; (2) la combinación de las anteriores formas para hombres y mujeres, así como la elaboración de nuevas normas para todas las escalas ocupacionales en nuevas muestras de hombres y mujeres, y (3) un incremento considerable en el número de escalas para ocupaciones técnicas cuyos requisitos de ingreso son inferiores al título universitario, una categoría subrepresentada en las ediciones anteriores.

El SII-Forma T317: Descripción general. El actual Inventario de Intereses de Strong consta de 317 reactivos agrupados en ocho partes. En las cinco primeras, el examinado registra sus preferencias marcando G, I o D para indicar "Gusto", "Indiferencia" y "Disgusto". Los reactivos en las primeras cinco partes caen en las siguientes categorías: ocupaciones, materias escolares, actividades (por ejemplo, dar una conferencia, reparar un reloj, recolectar dinero para caridad), actividades recreativas y contacto cotidiano con diferentes clases de personas (por ejemplo, ancianos, oficiales militares, gente que vive de manera peligrosa). Otras dos partes requieren que el examinado exprese sus preferencias entre pares de actividades (tratar con cosas en comparación con el trato con personas) y entre todas las posibles parejas de cuatro reactivos del mundo del trabajo (por ejemplo, ideas, datos, cosas y gente). Finalmente, una parte del inventario pide al individuo que responda a un conjunto de afirmaciones autodescriptivas marcando "Sí", "No" o "?".

El Strong se califica únicamente por computadora, en centros de calificación designados por la casa editora o mediante el uso de programas de cómputo que pueden adquirirse en una gran variedad de opciones. La figura 14.1 muestra la primera página del último perfil del Strong, que proporciona una "instantánea" o resumen de las puntuaciones más altas del examinado en las principales escalas del inventario.[6] Hay tres niveles de calificación, que difieren en su amplitud. Las más amplias y generales son las seis calificaciones de Temas Ocupacionales Generales; la siguiente subdivisión incluye 25 Escalas Básicas de Interés, y el nivel más específico proporciona las 211 Escalas Ocupacionales disponibles. Además, la Forma T317 del Strong genera calificaciones para cuatro nuevas Escalas de Estilo Personal que evalúan las preferencias en el Estilo de Trabajo, el Ambiente de Aprendizaje, el Estilo de Liderazgo y el Correr Riesgos y Aventuras. La forma del perfil también proporciona un conjunto de Índices Administrativos, incluyendo el número total de respuestas (para detectar un exceso de omisiones), el número de respuestas poco frecuentes o inusuales y el porcentaje de respuestas para las tres opciones en las ocho partes del inventario de Strong. Los índices pueden utilizarse como verificaciones de respuestas descuidadas o de conjuntos especiales de respuestas.

[6] El perfil completo del SII ha sido rediseñado y ahora consta de seis páginas. La carátula de cada página presenta las puntuaciones del individuo examinado en todas las escalas, y la parte trasera contiene información general acerca del significado de las puntuaciones y sugerencias sobre cómo proceder con la exploración de la carrera. Además del perfil, también pueden obtenerse informes narrativos más extensos y cualitativos con interpretaciones adaptadas a las puntuaciones del individuo.



INVENTARIO DE INTERESES DE STRONG

Informe del perfil para: CLIENTE 1
ID:
Edad: 20
Género: Masculino
Fecha de aplicación:
Fecha de calificación:

Página 1 de 6

INSTANTÁNEA: UN RESUMEN DE LOS RESULTADOS PARA CLIENTE 1

TEMAS OCUPACIONALES GENERALES

Los temas ocupacionales generales describen los intereses en seis áreas muy generales, que incluyen el interés en el trabajo y las actividades recreativas, clases de gente, y escenarios de trabajo. Sus intereses en cada área se muestran a la derecha de manera ordenada. Advierta que cada tema tiene un código, representado por la inicial de su nombre.

Usted puede emplear su código del tema, impreso debajo de los resultados, para identificar las materias escolares, trabajos de medio tiempo, licenciaturas, actividades recreativas o carreras que podrían resultarle interesantes. Vea la parte posterior de este perfil para encontrar sugerencias sobre cómo utilizar su código temático.

CÓDIGO TEMÁTICO	TEMA	MUY POCO INTERÉS	POCO INTERÉS	INTERÉS PROMEDIO	ALTO INTERÉS	MUY ALTO INTERÉS	INTERESES TÍPICOS
I	INVESTIGACIÓN	☐	☐	☑	☐	☐	Investigar, analizar
R	REALISTA	☐	☐	☑	☐	☐	Construir, reparar
C	CONVENCIONAL	☐	☐	☑	☐	☐	Explicar, procesar datos
A	ARTÍSTICO	☐	☐	☑	☐	☐	Crear o disfrutar el arte
S	SOCIAL	☐	☑	☐	☐	☐	Ayudar, enseñar
E	EMPRENDEDOR	☐	☑	☐	☐	☐	Vender, dirigir

Su código temático es IRC (vea la explicación a la izquierda).
Usted puede explorar ocupaciones con códigos que contengan cualquier combinación de esas letras.

ESCALA DE INTERESES BÁSICOS

Las escalas de intereses básicos miden su interés en 25 áreas o actividades específicas. A la derecha se presentan de manera ordenada sólo las cinco áreas en las que muestre el mayor interés. En la página 2 se presentan sus resultados en las 25 escalas de interés básicos.

A la izquierda de cada escala se encuentra una letra que muestra con cuál de los seis Temas Ocupacionales Generales está más relacionada. Esos códigos pueden resultarle de ayuda para identificar otras actividades que puede disfrutar.

CÓDIGO TEMÁTICO	INTERÉS BÁSICO	MUY POCO INTERÉS	POCO INTERÉS	INTERÉS PROMEDIO	ALTO INTERÉS	MUY ALTO INTERÉS	INTERESES TÍPICOS
R	ATLETISMO	☐	☐	☐	☐	☑	Jugar o mirar deportes
I	MATEMÁTICAS	☐	☐	☐	☐	☑	Trabajar con números o estadísticas
I	CIENCIAS MÉDICAS	☐	☐	☐	☑	☐	Trabajar en medicina o biología
R	ACTIVIDADES MECÁNICAS	☐	☐	☑	☐	☐	Trabajar con herramientas y equipo
A	ARTES APLICADAS	☐	☐	☑	☐	☐	Producir o disfrutar del arte visual

ESCALAS OCUPACIONALES

Las escalas ocupacionales miden qué tan similares son sus intereses a los de la gente que está satisfecha de trabajar en esas ocupaciones. A la derecha se presentan de manera ordenada únicamente las 10 escalas en las que sus intereses son más parecidos a los de esa gente. En las páginas 3, 4 y 5 puede encontrar sus resultados en las 211 escalas ocupacionales.

Las letras a la izquierda de cada escala identifican el tema o temas que describen más cercanamente los intereses de la gente que trabaja en esas ocupaciones. Puede usar las letras para encontrar otras ocupaciones relacionadas que tal vez le resulten interesantes. Después de revisar sus resultados en las seis páginas de este perfil, regrese a la página 5 para encontrar sugerencias sobre cómo hallar otras ocupaciones en el tema o temas que más le interesen.

CÓDIGO TEMÁTICO	OCUPACIÓN	MUY DIFERENTE	DIFERENTE	NIVEL MEDIO	SIMILAR	MUY SIMILAR
RI	INGENIERO	☐	☐	☐	☐	☑
RI	RADIOTÉCNICO	☐	☐	☐	☐	☑
IR	DENTISTA	☐	☐	☐	☑	☐
R	PLOMERO	☐	☐	☐	☑	☐
CI	ACTUARIO	☐	☐	☐	☑	☐
IR	OPTOMETRISTA	☐	☐	☐	☑	☐
CE	CONTADOR	☐	☐	☐	☑	☐
IR	QUÍMICO	☐	☐	☐	☑	☐
R	MECÁNICO AUTOMOTRIZ	☐	☐	☐	☑	☐
RIC	ELECTRICISTA	☐	☐	☐	☑	☐

***ESCALAS DE ESTILO PERSONAL** Miden sus niveles de comodidad en relación con el estilo de trabajo, el ambiente de aprendizaje, el estilo de liderazgo y los riesgos y las aventuras. Esta información puede resultarle de ayuda para tomar decisiones acerca de los ambientes particulares de trabajo y las clases de actividades que pueden serle satisfactorias. Sus resultados en esas cuatro escalas se encuentran en la página 6.*

CPP CONSULTING PSYCHOLOGISTS PRESS, INC. • 3803 Bayshore Road, Palo Alto, CA 94303

Figura 14.1. Resumen de resultados en el Inventario de Intereses de Strong. La primera página del perfil proporciona una "instantánea" de las puntuaciones obtenidas por un estudiante varón, universitario y de 20 años al elegir una licenciatura.

(Forma reproducida con autorización del editor de la p. 236 de *Strong Interest Inventory: Applications and technical guide*, de Lenore W. Harmon, Jo-Ida C. Hansen, Fred H. Borgen y Allen L. Hammer. Copyright © 1994 por Stanford University Press.)

La clasificación que hace el SII de los intereses ocupacionales proviene del modelo teórico desarrollado por John Holland (1966, 1985/1992) y se fundamenta en la extensa investigación realizada por este autor y por otros investigadores independientes. Los Temas Ocupacionales Generales identificados por el modelo de Holland se designan como Realista (R), de Investigación (I), Artístico (A), Social (S), Emprendedor (E) y Convencional (C).[7] Cada tema caracteriza no sólo un tipo de persona, sino también el ambiente de trabajo que encuentra más conveniente. Los entornos suelen estar poblados y dominados por personas del tipo correspondiente. De acuerdo con Holland, las personas no se clasifican rígidamente en uno de los seis tipos principales, sino que se caracterizan por el grado de parecido con uno o más tipos. Tales combinaciones, ordenadas por el grado de parecido, proporcionan una multiplicidad de patrones o "códigos" para describir la gran diversidad de las diferencias individuales.

La figura 14.2 presenta los seis temas en las esquinas del modelo hexagonal preparado por Holland. Para abreviar, se acostumbra utilizar las iniciales de cada uno de los temas para designarlos. Así, el orden de los temas en el hexágono forma las siglas R-I-A-S-E-C, que a su vez se emplea para designar al modelo. Debe notarse que las correlaciones más altas se obtuvieron entre las escalas que ocupan las posiciones adyacentes en el perímetro del hexágono y que las más bajas fueron las que resultaron entre las escalas en los extremos opuestos de las diagonales; por ejemplo, la Escala Realista correlaciona .53 con la Escala de Investigación, pero apenas .04 con la Artística y .06 con la Escala Social. De modo similar, si se grafican las Escalas Ocupacionales sobre el hexágono, casi todas siguen el orden esperado a lo largo del perímetro; por ejemplo, la Ingeniería, codificada como RI, cae entre Realista e Investigación mientras que Banquero, codificado CE, cae entre Convencional y Emprendedor. Comúnmente, las ocupaciones que obtienen una alta puntuación en un tema consiguen una baja en su opuesto directo (digamos, Artístico y Convencional). Cuando una Escala Ocupacional muestra correlaciones sustanciales con temas en los puntos opuestos del hexágono, a menudo incluye subgrupos heterogéneos con funciones ocupacionales claramente diferentes.

Las 25 *Escalas de Intereses Básicos* están clasificadas dentro de los seis Temas Ocupacionales Generales. Estas escalas constan de grupos de reactivos sustancialmente correlacionados. El contenido de las Escalas de Intereses Básicos es más homogéneo que el de las Escalas Ocupacionales, por lo que contribuyen a entender la razón de que un individuo califique alto en una Escala Ocupacional particular.

Las *Escalas Ocupacionales*, que constituyen el cuerpo principal del SVIB original, ahora se agrupan bajo los Temas Ocupacionales Generales apropiados. Al continuar el programa de investigación de este inventario se han agregado nuevas escalas y las anteriores han sido actualizadas con nuevas muestras de criterio. La Forma T317 incluye 211 Escalas Ocupacionales, de las cuales el 83 por ciento se normalizó con muestras obtenidas en las décadas de los ochenta y los noventa. Para todas esas escalas, salvo siete, se ha demostrado que es posible obtener suficientes individuos que

[7] Esos temas se traslapan considerablemente con las actitudes evaluativas medidas por el Estudio de Valores (Allport *et al.*, 1960) que, a su vez, fueron sugeridos por los *Tipos de Hombres* de Spranger (1928).

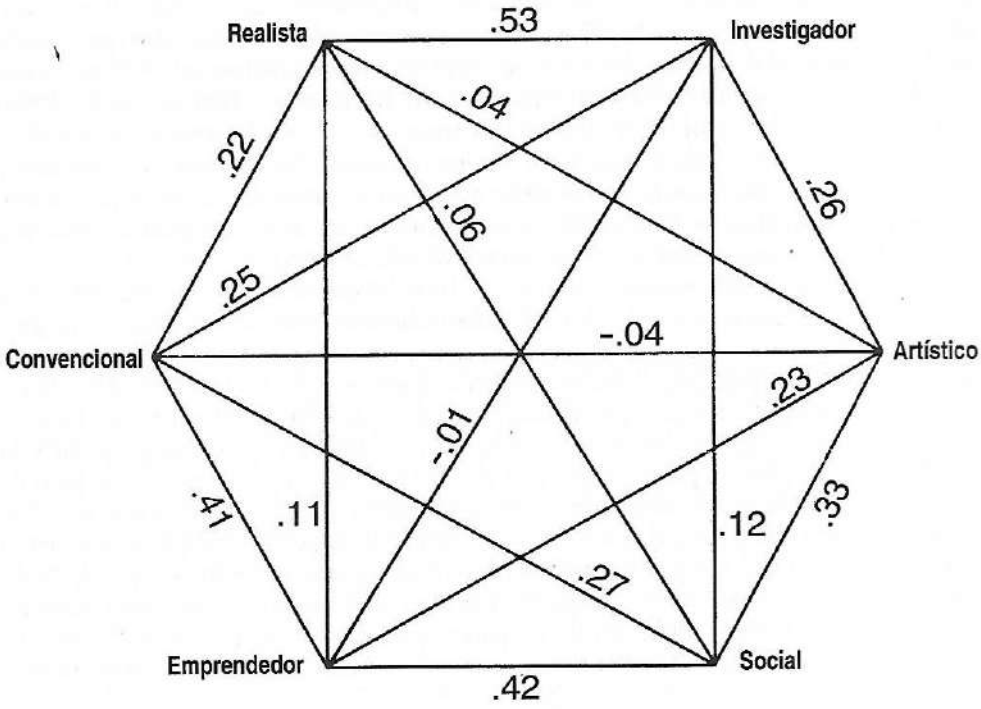


Figura 14.2. Modelo hexagonal de los Temas Ocupacionales Generales de Holland. Para cada par de temas, la correlación se basa en la Muestra General de Referencia formada por 9 484 hombres y 9 467 mujeres.

(Reproducido con autorización del editor de p. 51, *Strong Interest Inventory: Applications and Technical guide*, de Lenore W. Harmon, Jo-Ida C. Hansen, Fred, H. Borgen y Allen L. Hammer. Copyright © 1994 por Stanford University Press.)

permitan desarrollar escalas normalizadas para mujeres y para hombres (es decir, 102 escalas para cada género). Hasta ahora, cinco escalas han sido normalizadas únicamente con mujeres y dos sólo con hombres.

Casi todas las muestras empleadas al elaborar cada escala ocupacional contenían 200 personas o más, aunque el número real varía entre 60 y 1 187. En la revisión de 1994, se examinó a más de 55 000 personas, de las cuales poco menos de 40 000 cumplían con las especificaciones requeridas para utilizar en el desarrollo de la escala. Los grupos ocupacionales de criterio estaban formados principalmente por personas entre los 25 y los 60 años, empleadas en una determinada ocupación al menos durante tres años, que afirmaban sentirse satisfechos con su trabajo y que realizaban las tareas características de los miembros de esa ocupación. También se contó con dos muestras generales de referencia (MGR) que incluían a 9 484 hombres y 9 467 mujeres, los cuales se evaluaron en los años noventa. Las MGR comprenden 98 ocupaciones diferentes, de las que 90 están representadas por 200 miembros, seleccionados al azar de un banco disponible para cada ocupación; ocho ocupaciones se encuentran represen-

tadas por muestras de entre 92 y 195 miembros únicamente.[8] Los niveles de escolaridad de los individuos de las MGR en el Strong son mucho más elevados que para la población general de los Estados Unidos; en efecto, aproximadamente 80 por ciento de ellos tiene un grado de licenciatura, maestría o superior (Harmon, *et al.*, 1994, p. 110). Esta proporción de graduados universitarios supera las cifras comparables para la población general, incluso para la gente que tiene los trabajos profesionales y de oficina que constituyen la mayor parte de las ocupaciones en la MGR de Strong. Así, es posible criticar la última edición del Strong, al igual que sus predecesoras, por la falta de representatividad de sus muestras (Worthen, 1995).

Los reactivos del SII se asignaron a las escalas mediante dos métodos diferentes. Los del Tema Ocupacional General y de las Escalas Básicas de Intereses se reunieron en grupos homogéneos en función de las semejanzas identificadas por un análisis factorial de su contenido y las respuestas dadas por la gente. Por otro lado, los reactivos de cada escala ocupacional se seleccionaron y pesaron de acuerdo con las *diferencias* en los porcentajes reactivo-respuesta observadas entre la muestra de criterio ocupacional y la MGR para cada género; por ejemplo, en la escala para mujeres contadoras un peso de +1 indica que la respuesta ocurre más frecuentemente y una de –1 que ocurre menos a menudo entre las contadoras que entre las mujeres en general. Cualquiera que sea la frecuencia con la que fueran elegidas por estas profesionales, no se incluyeron en la escala correspondiente las respuestas que no lograron diferenciar en forma clara entre las contadoras y las mujeres en las MGR. La puntuación cruda total para cada escala ocupacional es la suma algebraica de los pesos positivos y negativos de los respondientes.

Calificación e interpretación. Todas las calificaciones del inventario de Strong se informan como puntuaciones estándares con una media de 50 y una *DE* de 10. La MGR de géneros combinados (N = 18 951) es la muestra normativa en que se calculan las calificaciones estándares tanto para los Temas Ocupacionales Generales como para las Escalas de Intereses Básicos. Aunque las etiquetas interpretativas y las representaciones gráficas de las puntuaciones en los perfiles se basan en las normas del mismo género, también pueden hacerse comparaciones con las normas del género opuesto, haciendo referencia a las gráficas presentadas en las formas para los perfiles. En las Escalas Ocupacionales cada individuo recibe dos calificaciones estándares, una derivada de la muestra ocupacional femenina y la otra de la muestra ocupacional masculina, lo que no sólo permite hacer comparaciones con el propio género, sino también obtener los datos necesarios para que consejeros e individuos hagan las comparaciones entre géneros para una mayor y mejor interpretación de los patrones de respuesta en los casos individuales.

La Guía Técnica de Aplicaciones del SII (*SII Applications and Technical Guide*, Harmon *et al.*, 1994) resulta de una sorprendente calidad y contiene mucha información que ayuda al consejero a interpretar los resultados y analizar con el cliente sus posibles implicaciones; por ejemplo, los perfiles que producen resultados consistentes internamente suelen tener una mayor predictibilidad. Por otro lado, ciertas inconsistencias, digamos, entre las Escalas de Intereses Básicos y las Escalas Ocupacionales, brindan información útil sobre la naturaleza y las fuentes de las preferencias expresa-

[8] En Harmon *et al.* (1994, apéndice A) puede encontrar un listado completo de todas las muestras ocupacionales empleadas en la revisión realizada en 1994 del SII.

das por los individuos. Desde otro punto de vista, los resultados obtenidos con el inventario pueden generalizarse a otras ocupaciones afines mediante vínculos que han sido establecidos con las ocupaciones que aparecen en la lista del *Dictionary of Occupational Titles* (DOT) preparado por el Departamento del Trabajo de los Estados Unidos (1991) y otros materiales de referencia similares (Gottfredson y Holland, 1989; Harmon *et al.*, 1994, apéndices A y B). La *Guía Técnica de Aplicaciones* también contiene varios capítulos dedicados al uso del inventario de Strong en poblaciones especiales, como los grupos culturalmente diversos y las personas con discapacidades. Además, hay muchas otras publicaciones diseñadas para ayudar a clientes y consejeros a comprender y utilizar los resultados del inventario de Strong (véase, por ejemplo, Borgen y Grutter, 1995; Hirsh, 1995; Prince, 1995).

Evaluación psicométrica. El inventario de Strong ha sido sometido a un programa continuo de investigación que ha generado muchos datos sobre su confiabilidad y validez (D. P. Campbell, 1971, 1977; Hansen y Campbell, 1985; Harmon *et al.*, 1994). Para las Escalas Ocupacionales, la media de la confiabilidad en el retest para una muestra de 191 adultos empleados, después de intervalos de entre tres y seis meses, fue de .90; para las Escalas de Intereses Básicos fue de .86, y para los Temas Ocupacionales Generales de .89. La estabilidad a largo plazo de las Escalas Ocupacionales para las ediciones anteriores del inventario de Strong también es elevada. Las correlaciones para periodos que se extienden por más de 20 años caen principalmente entre .60 y .70 para individuos menores de 25 años y en .80 para los mayores.

Los datos de validación concurrente los proporciona el grado de diferenciación entre varias muestras ocupacionales, y entre muestras ocupacionales y muestras de referencia. Para las 211 muestras ocupacionales incluidas en la revisión realizada en 1994, la media del traslape fue de 36 por ciento, lo que refleja una separación promedio de poco menos de dos desviaciones estándares entre las escalas ocupacionales y las MGR. La validez predictiva ha sido verificada en diversas muestras en intervalos largos para las versiones previas del inventario de Strong, y las evidencias indican una correspondencia considerable entre el perfil ocupacional inicial y la ocupación elegida a la postre. Un ejemplo concreto se encuentra en un seguimiento realizado durante 40 años de una muestra de psicólogos cuyas carreras profesionales revelaron algunas relaciones interesantes entre lo homogéneo o distintivo de sus perfiles originales y la frecuencia de cambios de puesto y cambios del trabajo en la enseñanza o en la investigación a tareas administrativas o trabajo aplicado (Vinitsky, 1973). En otra investigación se descubrió similitudes sorprendentemente altas en el perfil transcultural entre muestras de psicólogos evaluados en nueve naciones occidentales (Lonner y Adams, 1972). A pesar de la base extensa de investigación que los inventarios de Strong han acumulado a lo largo de los años y de su aparente solidez como indicadores de la elección y ejercicio ocupacional, es necesario emprender nuevos estudios que utilicen la última revisión para investigar directamente su validez predictiva.

En relación con la validación de constructo, resultan de particular importancia las relaciones de las Escalas Ocupacionales con los Temas Ocupacionales Generales, así como las interrelaciones que se ilustran en la figura 14.2. El modelo R-I-A-S-E-C establecido por Holland (1966, 1985/1992) ha tenido un enorme valor heurístico en el estudio de los intereses vocacionales y los inventarios de Strong han sido a su vez parte

integral de la tradición de investigación, junto con otras mediciones de intereses que emplean el mismo esquema. Por lo general, los descubrimientos empíricos realizados con el Strong y otros inventarios han sido consistentes con las predicciones del modelo.

La estructura y organización del SII han permitido su modificación y expansión en sus continuas revisiones. El inventario de Strong no sólo es pionero en el campo de la medición de intereses, sino que también es el de mayor uso entre los psicólogos consejeros de los Estados Unidos (Watkins, Campbell y Nieberding, 1994). Muy pocos instrumentos se han utilizado durante tanto tiempo y tan ampliamente; sin embargo, el tiempo transcurrido desde los sesenta ha sido testigo de un rápido crecimiento en el desarrollo de nuevos instrumentos en el área, lo que refleja la mayor atención a la exploración de profesiones y el reconocimiento de que los intereses desempeñan una función crucial en dicha exploración.

Habitualmente, los programas generales para la exploración de profesiones que veremos en el capítulo 17 incluyen medidas de intereses ocupacionales, que se usan en combinación con las calificaciones de tests de aptitudes múltiples y con la información laboral. Por ejemplo, los datos sobre los intereses obtenidos con el Inventario de Intereses Profesionales (*Career Interest Inventory*, CII) puede usarse junto con los Tests de Aptitudes Diferenciales con los que fue estandarizado (Psychological Corporation, 1991a). De igual modo, el Estudio de Aptitudes Ocupacionales (*Occupational Aptitude Survey*) y el Programa de Intereses 2 (*Interest Schedule-2* OASIS-2, Parker, 1991a, 1991b) se diseñaron para ayudar a los estudiantes de educación media en la exploración vocacional y el desarrollo profesional pues les proporcionan calificaciones en seis factores de aptitudes y 12 de intereses. Otra variante en el área de la planeación profesional es el Sistema Revisado de Toma de Decisiones Relacionadas con la Elección de Carrera de Harrington y O'Shea, (*Harrington-O'Shea Career Decision-Making System-Revised*, CDM-R, Harrington y O'Shea, 1993) que intenta integrar los datos autodescriptivos sobre intereses, valores y habilidades con la información laboral. El Estudio de Habilidades e Intereses de Campbell (*Campbell Interest and Skill Survey*, CISS, Campbell, Hyne y Nilsen, 1992) también mide los intereses y habilidades informados por el propio sujeto y se organiza de manera parecida al inventario de Strong, con quien David P. Campbell (autor del CIIS) trabajó durante mucho tiempo. La adición de los datos sobre habilidades permite hacer comparaciones de patrones altos y bajos de calificaciones en las escalas de intereses y habilidades, lo que a su vez extiende la base para la exploración profesional y la toma de decisiones proporcionada por la encuesta. En Kapes *et al.* (1994) pueden encontrarse revisiones del CII, el OASIS-2, el CDM-R y el CISS.

INVENTARIOS DE INTERESES: RECAPITULACIÓN Y ALGUNOS PUNTOS DESTACADOS

Entre los muchos inventarios de intereses de los que actualmente se dispone, revisaremos cuatro, cada uno de los cuales ilustra alguna característica digna de mención por su orientación teórica, su metodología o la población para el que fue diseñado; sin embargo, no trataremos de describirlos o evaluarlos en detalle porque todos fueron revisados hace poco en el *Mental Measurements Yearbook*, ya sea en su versión actual o en la anterior.

Estudio de Intereses Vocacionales de Jackson (*Jackson Vocational Interest Survey*, JVIS). Elegimos el JVIS (Jackson, 1977) en primer lugar porque ejemplifica el uso de complejos procedimientos de construcción de pruebas y, segundo, porque su planteamiento contrasta en varios aspectos con el seguido por el SII. Entre la publicación de ambos inventarios hay un lapso de 50 años. El Strong se concentra en ocupaciones específicas, tanto en la selección de reactivos como en las interpretaciones normativas, mientras que el Jackson utiliza grandes áreas de interés tanto en la preparación de reactivos como en el sistema de calificación. Las escalas ocupacionales de Strong son un ejemplo extremo de clave de criterio empírico y de validación relacionada con el criterio, mientras que el Jackson ejemplifica la validación de constructo en cada etapa de su desarrollo. En el Strong, el examinado responde independientemente a la mayor parte de sus reactivos marcando Gusto, Indiferencia o Disgusto, en tanto que todos los reactivos del Jackson son de elección forzada.

Como en el desarrollo de la Forma de Investigación de la Personalidad y el Inventario de Personalidad de Jackson que explicamos en el capítulo 13, el primer paso en la elaboración del JVIS consistió en definir los constructos o las dimensiones que pretendían medirse. Estas dimensiones, elegidas con base en la investigación publicada sobre la psicología del trabajo y las clasificaciones factoriales y lógicas de los reactivos de intereses vocacionales, eran de dos tipos, uno definido en términos de funciones laborales y el otro de estilos de trabajo. Las *funciones laborales* atañen a lo que hace el individuo en el empleo. Algunas están asociadas con una ocupación en particular, como la ingeniería, la jurisprudencia o la educación elemental; otras, como las relaciones humanas, la administración y el consejo profesional, se encuentran en muchas ocupaciones. Los *estilos de trabajo* no se refieren a actividades relacionadas con el empleo, sino a las preferencias por ambientes o situaciones de trabajo en las que se espera determinado comportamiento. En general, las dimensiones de los estilos de trabajo se relacionan con los valores del individuo, sea directa o indirectamente. Algunos ejemplos de estilos de trabajo comprenden la planeación, la independencia y el liderazgo dominante.

El desarrollo del inventario siguió varios pasos, incluyendo las evaluaciones y el análisis estadístico de los reactivos preparados para ajustarse a especificaciones detalladas de cada una de las funciones y los estilos. A partir de un banco inicial de más de 3 000 reactivos, que al principio sólo admitían una respuesta de gusto o disgusto, el procedimiento incluyó el análisis factorial de subconjuntos de reactivos preparados para cada escala. Antes de proseguir los análisis de reactivos se eliminaron estadísticamente los sesgos de respuesta, reflejados en un gran factor general. Se seleccionaron los reactivos que mostraban correlaciones altas con las calificaciones factoriales totales de su propia escala y correlaciones bajas con las otras escalas. Los pares de reactivos que representaban diferentes funciones laborales o estilos de trabajo y que habían mostrado frecuencias similares cuando se presentaron por separado se reunieron para el formato de elección forzada.[9]

La forma final del JVIS contiene 34 escalas de intereses básicos que cubren 26 funciones y ocho estilos de trabajo. El inventario se diseñó para aplicarse a ambos sexos, aunque también se dispone de normas percentilares separadas para subgrupos de

[9] Sin embargo, a diferencia de otros instrumentos de elección forzada, el emparejamiento de los reactivos del JVIS se realizó de tal manera que las calificaciones resultantes *no* son de naturaleza ipsativa.

mujeres y de hombres. Las normas se derivaron de grandes muestras de estudiantes de educación media y superior de los Estados Unidos y Canadá. Una alta puntuación en cualquiera de las 34 escalas de intereses básicos del JVIS indica un interés en las cosas que la gente hace en un campo particular de trabajo, así como en la forma en que se espera que actúe en cierto contexto laboral.

Las 34 escalas del JVIS pueden calificarse manualmente de manera rápida y conveniente, pero las opciones disponibles de calificación computarizada utilizan normas más recientes y proporcionan varios análisis de calificaciones adicionales en un informe breve o en uno más reciente, extenso y descriptivo, que brinda un material interpretativo y descriptivo individualizado así como gran cantidad de información que puede ayudar en la exploración profesional; por ejemplo, los informes generados por computadora incluyen calificaciones derivadas del análisis factorial de las 34 escalas de intereses básicos. Estas calificaciones, que siguen el modelo de los seis temas de Holland, incluyen los siguientes 10 Temas Ocupacionales Generales: Expresivo, Lógico, Inquisitivo, Práctico, Asertivo, Socializado, Servicial, Convencional, Emprendedor y Comunicativo. También se proporciona una medida de Satisfacción Académica, e Índices de Consistencia e Infrecuencia de Respuesta, así como un conteo de las respuestas no calificables. Otros análisis de las calificaciones incluyen la comparación del perfil total del individuo con los perfiles modales obtenidos por estudiantes universitarios en 17 grupos de los principales campos académicos y por participantes de 32 grupos ocupacionales. Los perfiles ocupacionales se derivaron de la aplicación conjunta de los inventarios de Jackson y de Strong que, mediante análisis estadísticos, permitió establecer una relación entre ambos, que a su vez proporcionó un medio de utilizar la extensa base de datos ocupacionales del Strong en la interpretación de las calificaciones del JVIS.[10]

Algunos revisores han sugerido que la redacción de los reactivos del JVIS puede resultar demasiado compleja para muchos estudiantes de secundaria (D. T. Brown, 1989; J. W. Shepard, 1989). Recientemente, se han hecho algunos cambios menores en el estilo del cuadernillo de reactivos, aunque se relacionan más con la actualización de la terminología que con su simplificación. También se anticipa la publicación de la revisión abreviada del manual que incorporará las normas obtenidas a mediados de los noventa; mientras tanto, es posible obtener un manual sobre las aplicaciones del JVIS en la consejería profesional (Verhoeve, 1993) y una guía ocupacional (Jackson, 1995).

El estudio de Intereses Ocupacionales de Kuder y sus predecesores. El inventario de intereses desarrollado por Frederic Kuder ha estado en uso casi tanto tiempo como la colección de Strong. El primero fue el Registro de Preferencias Vocacionales de Kuder (*Kuder Preference Record-Vocational*), cuya aproximación a la medición de los intereses difería de la de Strong en dos formas principales. Primera, Kuder utilizó tríadas de reactivos de elección forzada en las que los individuos indicaban cuál de tres actividades les gustaba más y cuál menos; segunda, no se obtenían calificaciones para ocupaciones particulares, sino para 10 áreas generales de interés (actividades al aire libre, mecánicas, de cómputo, científicas, persuasivas, artísticas, literarias, musicales, de servicio social y de oficina). Los reactivos para cada escala se formularon y

[10] Para una descripción completa de los procedimientos seguidos, véase Jackson (1977, capítulo 4) y Jackson y Williams (1975).

agruparon tentativamente sobre la base de la validez de contenido, mientras que la selección final se fundamentó en la consistencia interna y en las bajas correlaciones con otras escalas. Más adelante, se preparó el Estudio de Intereses Generales de Kuder (*Kuder General Interest Survey*, KGIS) como revisión y extensión hacia abajo del Registro de Preferencias Vocacionales de Kuder. Diseñada para usarse con alumnos de sexto a duodécimo grado, esta forma utiliza un lenguaje más simple y un vocabulario más sencillo que sólo requiere un nivel de lectura de sexto grado. Las revisiones del KGIS (M. Pope, 1995; D. Thompson, 1995) pueden encontrarse en la decimosegunda edición del MMY.

Una versión aún más reciente, el Estudio de Intereses Ocupacionales de Kuder (*Kuder Occupational Interest Survey*, KOIS) proporciona calificaciones que al igual que el inventario de Strong hacen referencia a grupos ocupacionales específicos (Kuder, 1966; Kuder y Diamond, 1979; Kuder y Zytowski, 1991); pero a diferencia del Strong, el KOIS no utiliza un grupo general de referencia, sino que la puntuación del individuo en cada escala ocupacional se expresa como correlación entre su patrón de intereses y el patrón de intereses del grupo ocupacional particular.[11] Este estudio de intereses puede calificarse en el lugar o puede hacerlo el editor mediante un escáner óptico, aunque también puede aplicarse y calificarse por computadora. Se dispone de calificaciones para 109 grupos ocupacionales concretos y para 40 licenciaturas universitarias. Algunas escalas se desarrollaron únicamente con hombres y otras sólo con mujeres, pero las puntuaciones de todas se informan para individuos de ambos sexos. El nivel de las ocupaciones cubiertas por el inventario es muy variable, y va del cosmetólogo y el conductor de camiones al químico y el abogado.

Mediante el análisis estadístico intensivo de las puntuaciones de 3 000 personas (100 en cada uno de los 30 grupos nucleares representativos de los campos de las ocupaciones y las licenciaturas cubiertos por el inventario), Kuder demostró que puede lograrse una mejor diferenciación entre los grupos ocupacionales con el sistema de calificación empleado en su estudio que con las escalas ocupacionales derivadas mediante el uso de un grupo general de referencia. Aún continúa la investigación sobre otras escalas ocupacionales (Zytowski, 1992).

El KOIS proporciona ahora puntuaciones ocupacionales y puntuaciones de 10 intereses básicos homogéneos y generales, denominados *Estimaciones de Intereses Vocacionales* (EIV), que son calificaciones percentilares derivadas de escalas breves equivalentes a las 10 calificaciones del área de interés del anterior Registro de Preferencias de Kuder. También pueden traducirse a los códigos temáticos R-I-A-S-E-C de Holland mediante la correspondencia directa de algunas escalas y promediando los percentiles de dos o tres escalas Kuder para el resto (por ejemplo, el promedio de las áreas Artística, Literaria y Musical de Kuder por la Artística de Holland). Por lo general, los críticos aprecian las características técnicas del KOIS a la vez que señalan su carencia de validez predictiva y su incapacidad para dirigir los efectos de su formato de elección forzada sobre las puntuaciones (véase Herr, 1989; Tenopyr, 1989).

[11] La correlación empleada es el coeficiente Lambda establecido por Clemans (1958). Se trata esencialmente de una correlación de punto biserial ajustada a las diferencias en la homogeneidad de diferentes grupos de criterio. La variable dicotómica consiste en las respuestas marcadas contra las no marcadas en la hoja de respuestas del individuo; la variable continua es la proporción de personas del grupo de criterio que marcaron cada respuesta.

Inventario de Evaluación de Carreras, Versión Vocacional (*Carrer Assessment Inventory-The Vocational Version*, CAI-VV.)[12] El CAI (Johansson, 1984), que se puso en circulación en 1975, sigue de cerca el patrón del inventario de Strong, pero en contraste con la mayor parte de los inventarios de intereses, se diseñó en particular para quienes buscan una carrera que *no* requiera una licenciatura de cuatro años o capacitación profesional avanzada. Se concentra en las habilidades para el comercio, el trabajo técnico y las ocupaciones de servicio. Algunos ejemplos de las escalas ocupacionales disponibles son: mecánico aeronáutico, técnico dental, trabajador de cafetería, programador de computadoras y enfermera registrada. Los 305 reactivos del inventario se agrupan en tres categorías de contenido: Actividades, Materias Escolares y Ocupaciones. Cada reactivo proporciona cinco alternativas de respuesta que van de "Me gusta mucho" a "Me disgusta mucho". Redactado a un nivel de lectura de sexto grado, el CAI también puede utilizarse con adultos que tienen pocas habilidades de lectura. Igual que el Strong, el CAI proporciona puntuaciones en tres escalas principales que comprenden las seis escalas de Temas Generales de Holland, 22 escalas homogéneas de Intereses Básicos y 91 escalas Ocupacionales; también se incluyen los Índices de Aplicación y cuatro escalas No Ocupacionales.

Aunque los procedimientos seguidos al desarrollar el CAI fueron muy similares a los del Strong, la obtención de los datos y los análisis estadísticos se realizaron independientemente para este inventario. A excepción de las escalas de Temas Generales, las escalas particulares elaboradas en cada categoría son, por consecuencia, específicas del CAI. El lector interesado en el CAI-VV puede encontrar una revisión en los trabajos de Kehoe (1992) y Vacc (1992).

Investigación Autodirigida (SDS). Otra aproximación a la evaluación de los intereses ocupacionales lo ejemplifica la Investigación Autodirigida (*Self-Directed Search*, SDS), instrumento preparado por Holland y cuyo modelo hexagonal de temas ocupacionales generales (que ya hemos visto) atrajo gran atención y ha sido incorporado a varios de los inventarios actuales (Holland, 1985/1992; Holland, Fritzsche y Powell, 1994; Holland y Gottfredson, 1976; Holland, Powell y Fritzsche, 1994).

Como su nombre lo indica, la SDS se diseñó como instrumento de consejería vocacional aplicado, calificado e interpretado por el propio individuo. Aunque se organiza alrededor de los intereses, el procedimiento también requiere que se hagan autovaloraciones de habilidades y competencias informadas. El individuo llena el cuadernillo de autoevaluación, califica las respuestas y calcula seis puntuaciones resumidas que corresponden a los temas del modelo de Holland (Realista, Investigador, Artístico, Social, Emprendedor y Convencional). Las tres puntuaciones más altas se utilizan para encontrar un código de tres letras.[13] Posteriormente, se utiliza un cuader-

[12] Actualmente se dispone de dos versiones del CAI, la Versión Vocacional (VV) y la Versión Mejorada (VM). Nuestra descripción de esta sección sólo se refiere a la VV. La VM, aunque es de una estructura muy similar, es un instrumento completamente separado (Johansson, 1986) que puede aplicarse a un mayor número y gama de ocupaciones, incluyendo muchas que requieren educación posterior a la secundaria.

[13] Aunque la SDS se diseñó para ser autocalificado, el manual recomienda cierta supervisión y verificación de las puntuaciones. Un estudio de 107 individuos de diversas edades seleccionados al azar que presentaron la edición actual de la SDS mostraron que 7.5 por ciento había llegado a códigos que contenían letras o transposiciones incorrectas (Holland, Powell y Fritzsche, 1994, p. 16).

nillo adjunto, el localizador de ocupaciones, para encontrar entre 1 335 ocupaciones aquéllas cuyos códigos se parecen al código resumido del individuo. Éstas se eligieron de modo que representaran a casi todos los trabajadores de los Estados Unidos. Con un código de conversión, el usuario puede también explorar todas las ocupaciones incluidas en el *Dictionary of Occupational Titles*, además de que se le proporcionan instrucciones, procedimientos y fuentes adicionales de información para facilitarle las decisiones relacionadas con la elección de carrera.

La SDS se utiliza mucho en diversos medios y ha generado una gran cantidad de investigación, no sólo de parte de su autor, sino también de investigadores independientes. En los años transcurridos desde su publicación, ha sido sometida a diversas revisiones para simplificar el procedimiento y reducir el sesgo del género en las decisiones relacionadas con la elección de carrera. El principal atractivo se encuentra en su brevedad y simplicidad, en la característica de "hágalo usted mismo" y en su contribución a la expansión de las opciones profesionales del individuo. Además de la forma regular del inventario (Forma R), se cuenta con otras tres versiones del instrumento: (1) la forma F (Fácil), desarrollada para individuos con limitadas habilidades de lectura; (2) la Forma PC (Planeación de Carrera) diseñada para adultos que se encuentran en medio de transiciones profesionales, y (3) la versión de Exploración de Carreras, destinada para estudiantes de educación media y media superior.

Con respecto a las propiedades psicométricas de la SDS, los índices de confiabilidad son en general satisfactorios para las calificaciones resumidas. La validación de constructo de los seis temas básicos descansa sobre todo en la investigación que condujo a la formulación de los temas y en los subsecuentes análisis factoriales confirmatorios que en su mayor parte son favorables (véase, por ejemplo, Oosterveld, 1994). La validez concurrente y la eficacia predictiva de la SDS fluctúa de acuerdo con la estructura de las muestras en términos de la edad, género, nivel educativo y tipos de distribución (Holland, Fritzche y Powell, 1994). Las críticas al instrumento se centran en algunos de sus procedimientos de calificación e interpretación (M. H. Daniels, 1989; Manuele-Adkins, 1989). No obstante, los revisores coinciden en que el instrumento proporciona una forma sencilla, barata y relativamente precisa para explorar las vocaciones. Los lectores interesados disponen de una amplia bibliografía relativa a la SDS y varios cuadernillos informativos que pueden usar junto con la prueba.

J. L. Holland (1966, 1985/1992) concuerda con quienes consideran que las preferencias ocupacionales constituyen la elección de una forma de vida en que se reflejan el autoconcepto y las principales características de personalidad del individuo. Cada uno de los temas ocupacionales de Holland corresponden al "tipo" o grupo de atributos de personalidad, un individuo determinado puede ser descrito en términos de uno o más tipos predominantes. Los temas también corresponden a modelos ambientales en términos de cuáles son los diferentes ambientes ocupacionales que pueden caracterizarse. Estos ambientes no sólo comprenden los rasgos físicos y las demandas de trabajo, sino también las clases de personas con las que labora el individuo (colaboradores, supervisores, clientes, estudiantes). De acuerdo con Holland, las personas buscan ambientes que sean consistentes con su tipo de personalidad, congruencia que aumenta la satisfacción en el trabajo, el éxito y la estabilidad en el empleo.

Desde otro punto de vista, el acercamiento general de Holland a la evaluación de los intereses vocacionales coincide con ciertos desarrollos, respecto a la elec-

ción de carrera, en la psicología de la toma de decisiones. Super (1953, 1957, 1990) ha reiterado que las elecciones vocacionales implican poner en práctica el autoconcepto. Existe una larga tradición de investigaciones que aborda las diferencias de personalidad entre grupos ocupacionales (véase, por ejemplo, Borgen, 1986; Costa, McCrae y Holland, 1984; Osipow, 1973, capítulo 6; Pietrofesa y Splete, 1975, capítulo 4; Super y Bohn, 1970, capítulo 5). La elección de una ocupación a menudo refleja las necesidades emocionales básicas del individuo, y el ajuste ocupacional es un componente importante del ajuste general a la vida (Tait, Padgett y Baldwin, 1989). La evaluación de los intereses vocacionales y, más específicamente, la identificación de grupos ocupacionales cuyos intereses y actitudes comparte más el individuo, se convierten de esa forma en el punto central en la comprensión de personalidades diferentes.[14]

ALGUNAS TENDENCIAS SIGNIFICATIVAS

Desarrollo y uso del inventario. Entre los desarrollos más evidentes en las mediciones actuales de los intereses está la combinación de las dos principales posiciones teóricas en la psicología vocacional, así como de las diferentes aproximaciones en la elaboración de inventarios y la utilización cruzada de bancos de datos empíricos con propósitos de interpretación. Cada vez son más los instrumentos que proporcionan puntuaciones *tanto* en las escalas generales, homogéneas, de intereses *como* en las escalas ocupacionales específicas. Los seis temas ocupacionales del modelo de Holland han sido incorporados a un número creciente de inventarios recientes o revisados.

Desde otro ángulo, actualmente se establecen nexos con los datos ocupacionales empíricos de grandes poblaciones a los que se tiene acceso, lo que se ilustra en el uso de los datos ocupacionales del inventario de Strong en la interpretación de los resultados del Estudio de Intereses Vocacionales de Jackson; pero también en la afinidad de muchos inventarios actuales, como la SDS de Holland y el SII, con los datos proporcionados por el *Dictionary of Occupational Titles* (*U. S. Department of Labor Employment and Training Administration*, 1991). Existen desarrollos prometedores que fortalecen y aumentan la utilidad de cualquier instrumento. Los efectos deberían ser benéficos, toda vez que los nexos se establecen mediante procedimientos psicométricos adecuados y se toman las medidas necesarias para impedir la posible sobregeneralización en la interpretación de las puntuaciones.

Otra característica de los inventarios de intereses recientes o revisados es el aumento de los niveles ocupacionales que cubren. Al inicio, los inventarios de intereses se concentraron en las carreras profesionales y mostraban un conocimiento superficial de las ocupaciones que no requerían una licenciatura o la asistencia a una escuela de educación profesional. Aunque al principio se hicieron algunos esfuerzos por extenderlos hacia abajo (por ejemplo, Clark, 1961), los inventarios resultantes no eran de uso común. En contraste, ahora encontramos que la más reciente revisión del inventario de Strong ofrece un número considerable de escalas para ocupaciones vocacio-

[14] Waller, Lykken y Tellegen (1995) informan de ciertos descubrimientos pertinentes para las correlaciones entre los intereses ocupacionales, los intereses en el tiempo libre y los rasgos de personalidad en una muestra grande de participantes en el Registro de Gemelos de Minnesota. Sus estudios plantean que, aunque es claro que los rasgos de personalidad median el ajuste entre persona y ocupación, los tres conjuntos de características pueden conceptuarse como dominios diferentes.

nales o técnicas que no requieren de una licenciatura. Más aún, algunos de los instrumentos más recientes, como la versión vocacional del Inventario de Evaluación de Carreras (*Career Assessment Inventory*, Johansson, 1984) y el Inventario de Orientación Vocacional (*Career Directions Inventory*, Jackson, 1986), se diseñaron principalmente para personas cuya educación no va más allá de las escuelas secundaria o técnica. Es probable que esos desarrollos reflejen, al menos en parte, un mayor reconocimiento de la importancia que la elección adecuada de una carrera tiene en todos los niveles educativos, así como la función crucial que desempeñan los intereses en el éxito y la experiencia de una satisfacción personal con el trabajo en cualquier empleo.

Otra tendencia discernible tiene implicaciones que van más allá de los inventarios de intereses a otras formas de evaluación. En una provocadora mirada al futuro de los inventarios de intereses, J. L. Holland (1986) se refirió al mayor reconocimiento de tales inventarios como técnicas de intervención. ¿Qué *efectos* tiene el inventario sobre el examinado?; por ejemplo, puede favorecer y fortalecer las aspiraciones vocacionales que ya tiene, pero también puede estimularlo a explorar de manera más detallada el mundo del trabajo y a que preste atención a opciones que no había considerado; además, puede aumentar la comprensión que tiene de sí mismo. Esta diversidad de efectos puede reflejarse no sólo en las interpretaciones individualizadas de las puntuaciones, sino también en la aplicación y construcción de los inventarios de intereses, lo que a su vez puede representar otra prometedora aplicación de la evaluación adaptativa computarizada.

En la década de los noventa, la psicología de la elección vocacional y desarrollo de carreras se ha enriquecido por la aplicación de paradigmas en el campo de la psicología cognoscitiva (véase, por ejemplo, Peterson, Sampson y Reardon, 1991). La consideración de la toma de decisiones vocacionales como una actividad de solución de problemas, que tal vez recurra en el ciclo vital, ha subrayado la necesidad de brindar a los individuos el conocimiento y las habilidades de procesamiento de información que requieren para llegar a soluciones óptimas. Muchos de los nuevos instrumentos incluyen programas para la exploración de carreras que constituyen la extensión lógica de las opiniones de Super sobre el desarrollo profesional y del acercamiento de Holland a la elección vocacional. Se ha hecho cada vez más claro que el área de la conducta vocacional ofrece oportunidades sin igual para hacer que los individuos participen activa y directamente en el uso de los *insights* y las herramientas de la psicología (Borgen, 1991).

Modelos de ocupaciones. Casi todas las pruebas vocacionales que analizamos en este capítulo han surgido del modelo de temas ocupacionales de Holland (por ejemplo, el CAI y la SDS) o lo han incorporado a sus procedimientos en mayor o menor grado (por ejemplo, el SII o el KOIS). En efecto, durante las dos últimas décadas los planteamientos teóricos de Holland han estimulado el trabajo en el campo de la psicología vocacional, dentro y fuera de los Estados Unidos (véase, por ejemplo, Borgen, 1991; Lokan y Taylor, 1986). En una evaluación inusualmente comprensiva del modelo, Tracey y Rounds (1993) hicieron un meta-análisis estructural de las escalas R-I-A-S-E-C con diversos instrumentos. En este estudio, en el que se utilizó 104 muestras estadounidenses se descubrió un buen apoyo para el modelo; sin embargo, el ajuste no fue tan bueno en un meta-análisis similar realizado después por los mismos autores con muestras internacionales y de minorías estadounidenses (Rounds y Tracey, 1996). Otras investigaciones transnacionales y transétnicas han producido resultados contradictorios (Fouad y

Dancer, 1992; Hansen, 1987; Khan, Alvi, Shaukat y Hussain, 1990; Swanson, 1992). No resulta sorprendente que la investigación transcultural con el modelo R-I-A-S-E-C proponga que tal vez no sea igualmente aplicable entre culturas en todos sus aspectos.

En cualquier caso, la simplicidad del hexágono R-I-A-S-E-C y la considerable influencia que durante los últimos 20 años ha ejercido en la investigación ha llevado a muchos a plantear que puede ser ya tiempo de que el campo avance o, al menos, de agregar nuevos elementos que aumenten su utilidad.[15] Otros han apelado al desarrollo de nuevas estructuras teóricas que expliquen otros aspectos de las ocupaciones, como las capacidades y preferencias reforzantes, mientras que a otros más les gustaría ver que se investigaran variables y dimensiones adicionales (véase, por ejemplo, Dawis, 1992; Prediger y Vansickle, 1992; R. H. Schwartz, 1992). El propio Holland ha afirmado que la investigación ya no debería dirigirse a tratar de replicar el hexágono con muestras diferentes, sino al estudio de los usos e interpretaciones de los inventarios de intereses y a los efectos que éstos tienen sobre los individuos que los responden (Holland y Gottfredson, 1992).

Tracey y Rounds (1996) han realizado uno de los esfuerzos más ambiciosos por modificar y extender el modelo de Holland mediante el uso de nuevas y más complejas técnicas de análisis de datos, que incluyen el escalamiento multidimensional. Su representación tridimensional de los intereses vocacionales emplea la tipología de Holland y las dos dimensiones que supuestamente la fundan[16] como punto de partida para un modelo esférico que pueda dar cabida a un número flexible de intereses, dependiendo de las necesidades del examinado, así como una dimensión de *prestigio*. Un número reciente del *Journal of Vocational Behavior* se dedica especialmente a la presentación de este nuevo modelo y a los comentarios que sobre el mismo hacen diversos investigadores. En general, las respuestas al trabajo de Tracey y Round son muy positivas y muestran que quizá sirva como catalizador para la integración y el avance en el campo de la teoría y la medición vocacional (véase Borgen y Donnay, 1996; González, 1996; G. D. Gottfredson, 1996; Hansen, 1996; Harmon, 1996; Prediger, 1996).

La investigación sobre la naturaleza, el número y la organización de los intereses básicos iguala a la conducida para identificar los factores principales en el campo de las habilidades y la personalidad (revisada en los capítulos 11 y 13).[17] En los tres casos, las categorías descubiertas por el análisis de los datos dependen de las variables específicas y las muestras empleadas; más aún, esas categorías son entidades descriptivas, más que explicatorias, cuyo valor principal consiste en su capacidad para simplificar la obtención y el uso de información para evaluar y predecir el comportamiento.

ENCUESTAS DE OPINIÓN Y ESCALAS DE ACTITUD

Naturaleza de los instrumentos. A menudo se define la actitud como la tendencia a reaccionar de manera favorable o desfavorable a cierta clase de estímulos, co-

[15] Un número especial del *Journal of Vocational Behavior* (abril 1992), que es la fuente de varias de las referencias citadas en esta sección, se dedicó por completo a la consideración de la teoría de Holland.

[16] Esas dimensiones bipolares, identificadas por Prediger (1982) y ampliamente utilizadas desde entonces, son Personas/Cosas y Datos/Ideas.

[17] Véase en Rounds (1995) para una exposición esclarecedora y de gran alcance sobre los intentos históricos por identificar y clasificar las dimensiones de los intereses y para explorar sus correlaciones.

mo los grupos étnicos o nacionales, una costumbre o una institución. Es evidente que, cuando se las define de esa manera, las actitudes no pueden observarse directamente, sino que tienen que inferirse de la conducta abierta, verbal y no verbal. En términos más objetivos, puede decirse que el concepto de actitud implica *la consistencia de la respuesta* a ciertas categorías de estímulos. En la práctica, el término "actitud" ha sido más asociado con estímulos sociales y con respuestas con matices emocionales, y a menudo incluye también juicios de valor.

En ocasiones se distingue las opiniones de las actitudes, pero las diferenciaciones propuestas no son congruentes ni defendibles. Es más común que ambos términos se utilicen de manera indiferente y así los emplearemos en este análisis; sin embargo, en lo que respecta a la metodología de evaluación, tradicionalmente se distingue a las encuestas de opinión de las escalas de actitud. Por lo general, las *encuestas de opinión* implican respuestas a preguntas específicas que no necesitan ser relacionadas; las respuestas a dichas preguntas se mantienen separadas en lugar de combinarlas en una calificación total; por ejemplo, una encuesta de opinión aplicada a empleados podría incluir preguntas sobre los programas de trabajo, la forma de pago, los beneficios adicionales, la cafetería de la compañía y las relaciones con los supervisores; cada uno de esos reactivos se incluye por su importancia intrínseca para el mejoramiento de las relaciones con los empleados, por lo que las respuestas a cada pregunta se tabulan de manera separada para identificar las fuentes de satisfacción e insatisfacción de los trabajadores.[18]

Por otro lado, las *escalas de actitud* por lo general arrojan una calificación total que indica la dirección e intensidad de la actitud del individuo hacia una compañía, un grupo de gente, una política u otras categorías de estímulos. En la construcción de una escala de actitud, las diferentes preguntas se diseñan para medir una sola actitud o variable unidimensional, y por lo común se utilizan procedimientos objetivos en un esfuerzo por acercarse a esa meta; por ejemplo, una escala de actitud para empleados produce una sola calificación que muestra el grado de satisfacción con el trabajo o la actitud total del individuo hacia la compañía.

Principales tipos de escalas de actitud. En todas las escalas de actitud, los individuos indican su acuerdo o desacuerdo con una serie de afirmaciones sobre el objeto de la actitud. Se han elaborado procedimientos especiales para alcanzar la unidimensionalidad u homogeneidad de los reactivos, la igualdad de las distancias entre las unidades de la escala y la comparabilidad de las calificaciones de diferentes escalas. Se ha concedido gran atención a los problemas técnicos incluidos en la elaboración de las escalas de actitud, por lo que la metodología ha hecho notables avances teóricos y estadísticos. La revisión de las técnicas especializadas de escalamiento, que ahora constituyen un área floreciente de los métodos estadísticos, rebasa el alcance de este texto (Jones y Koehly, 1993; D. J. Mueller, 1986; Ostrom, Bond, Krosnick y Sedikides, 1994; Procter, 1993; Reckase, 1990; Young, 1984), sin embargo, podemos examinar brevemente los tres principales acercamientos a la construcción de escalas que es común encontrar en la bibliografía sobre la evaluación psicológica (estas aproximaciones son representadas por las escalas de tipo Thurstone, Guttman y Likert).

[18] Una serie de volúmenes editada por Fink (1995) proporciona una guía general dedicada especialmente a la planeación, preparación, implantación y análisis de encuestas.

La adaptación que hizo Thurstone de los métodos psicofísicos a la cuantificación de los datos obtenidos de jueces representó un hito en la construcción de escalas de actitud (Thurstone, 1959; Thurstone y Chave, 1929). Con esos procedimientos, Thurstone y sus colaboradores prepararon alrededor de 20 escalas para medir las actitudes hacia la guerra, la pena capital, la iglesia, el patriotismo, la censura y muchas otras instituciones, prácticas, temas y grupos nacionales o étnicos. La elaboración de una *escala tipo Thurstone* empieza con el acopio de una gran cantidad de afirmaciones que expresan una amplia gama de actitudes hacia el objeto considerado. Se solicita a numerosos jueces que clasifiquen las afirmaciones en pilas (usualmente 11) de acuerdo con el grado en que resultan favorables. Los jueces no indican sus propias actitudes, sino que se concretan a clasificar las afirmaciones. El valor que cada declaración tiene en la escala está determinado por la posición media que le asignaron. La variabilidad de los juicios se considera como indicador de la ambigüedad en la medida en que diferentes jueces asignen la misma afirmación a categorías distintas. Se elige los reactivos que muestran una variabilidad mínima y una gran extensión de valores de escala, y que se aproximan a un espaciamiento igual en el intervalo de 11 puntos. En la escala final de actitud, las afirmaciones se presentan en orden aleatorio sin indicaciones de sus valores de escala, y la puntuación que recibe el individuo es la media del valor escala de todas las afirmaciones que aprueba.

La *escala tipo Guttman* se desarrolló originalmente como una técnica para determinar si un conjunto de afirmaciones de actitud era unidimensional (L. Guttman, 1944, 1947). Según Guttman, la escala es perfecta si el individuo que concuerda con cierta afirmación de una actitud particular también lo hace con afirmaciones más moderadas de esa actitud. En otras palabras, los reactivos de esa escala de actitud pueden ordenarse en un continuo de intensidad o dificultad de aceptación, por lo que la posición de cada persona en la escala determinaría por completo sus respuestas. Se supone que si conocemos cuáles son las afirmaciones más extremas que el individuo puede aceptar, deberíamos ser capaces de reproducir todas sus respuestas; sin embargo, debido a los errores de medición en cada respuesta, en la práctica sólo es posible aproximarse dentro de ciertos límites. El procedimiento esencial en el desarrollo de una escala Guttman consiste en identificar un conjunto de reactivos que caigan en una secuencia ordenada en términos de la aprobación que reciben de los individuos que las contestan. Los reactivos que no cumplen con este requisito son eliminados. Cabe recordar que el mismo concepto de ordinalidad o progresión uniforme de la ejecución es la base de las escalas piagetianas que vimos en los capítulos 3 y 9.

Como la construcción de una escala Thurstone requiere procedimientos más bien complicados, y dado que es difícil cumplir en la práctica las condiciones de la escala Guttman, Likert (1932) desarrolló una escala que es más sencilla de construir y cuya confiabilidad es igualmente satisfactoria. La *escala tipo Likert* comienza con una serie de afirmaciones, cada una de las cuales expresa una actitud favorable o desfavorable. Los reactivos se seleccionan en función de las respuestas de personas a las que se les aplicó el proceso de construcción del instrumento. La base principal para la selección de los reactivos es la consistencia interna, aunque también se utilizan criterios externos cuando están disponibles. Las escalas Likert requieren una respuesta graduada en cada afirmación, que se expresa en términos de las siguientes cinco categorías: total-

mente de acuerdo (TA), de acuerdo (A), indeciso (I), en desacuerdo (D), totalmente en desacuerdo (TD). Para calificar la escala, las opciones de respuesta se acreditan como 5, 4, 3, 2 o 1 del extremo favorable al desfavorable. Por ejemplo, estar "totalmente de acuerdo" con una afirmación favorable recibiría una puntuación de cinco, igual que estar "totalmente en desacuerdo" con una afirmación desfavorable. La suma de las calificaciones de los reactivos representa la puntuación total del individuo, que debe interpretarse en términos de normas establecidas empíricamente.

Casi todas las escalas de actitud se han destinado a proyectos de investigación particulares; algunas se diseñaron para la investigación de las actitudes y la moral de los empleados, otras han sido usadas para evaluar los resultados de programas educativos y de capacitación. Las escalas de actitud pueden contribuir a la evaluación de diferentes procedimientos instruccionales diseñados para modificar ciertas actitudes particulares, o pueden utilizarse después de determinado programa educativo para medir los cambios en las actitudes de los estudiantes hacia la literatura, el arte, diferentes grupos étnicos y culturales o los problemas sociales y económicos. Una de las aplicaciones más difundidas de la medición de las actitudes se encuentra en la investigación de la psicología social: prácticamente cualquier libro de texto sobre la materia incluye secciones sobre las actitudes y su medición. Entre los muchos problemas investigados con la medición de las actitudes podemos mencionar las diferencias grupales en las actitudes, la función que éstas cumplen en las relaciones entre grupos, los factores antecedentes asociados con su desarrollo, la interrelación de las actitudes (incluyendo análisis factoriales y otros métodos de análisis multivariados), tendencias y cambios temporales en las actitudes y la modificación experimental de las actitudes mediante experiencias intercaladas.[19] Son relativamente pocas las escalas de actitud que se han publicado, aunque la mayor parte se describe por completo en la bibliografía correspondiente. M. E. Shaw y Wright (1967) recopilaron en un libro una extensa colección de las primeras escalas de actitud construidas para diversos propósitos, y puede encontrarse información sobre las mediciones más recientes de ciertas actitudes y valores, como la alienación y la anomía, la autoestima y el locus de control en *Measures of Personality and Social Psychological Attitudes*, compilado por Robinson, Shaver y Wrightsman (1991).

Una nota sobre las mediciones y las variables relacionadas con el género. Las ediciones anteriores de este libro incluían secciones que cubrían con cierta profundidad las medidas de los roles sexuales y conceptos relacionados como masculinidad, femineidad y androginia, lo que no sucede en esta edición. Aunque la investigación y los instrumentos que pretenden evaluar esos fenómenos aún proliferan (véase, por ejemplo, Lenney, 1991), muchos de los estudiosos del campo están de acuerdo en que se encuentran en un estado de desorden conceptual. Desde el punto de vista más general de las diferencias individuales en las variables relacionadas con el género, Betz (1995) preparó una excelente actualización del estado de los conocimientos relacionados con estas variables y llegó a la conclusión de que la carencia de un marco teórico y

[19] En Eagly y Chaiken (1993) puede encontrarse una excelente compilación de la vasta especialidad de la psicología social que estudia las actitudes. Para una revisión breve de la bibliografía sobre las actitudes y el cambio de actitud, véase Olson y Zanna (1993).

de definiciones conceptuales sustantivas ocasionaron que se limitara el progreso en el estudio de los fenómenos relacionados con el género. A lo anterior debería sin duda añadirse que las dos últimas décadas también han presenciado cambios sin precedentes en las concepciones culturales del género en todo el mundo. En cualquier caso, como Betz señala, los conceptos globales que pretenden explicar las diferencias relacionadas con el género, como los roles sexuales o masculinidad-femineidad, no han recibido apoyo de los descubrimientos empíricos. Según ella, para que pueda ocurrir un mayor progreso es necesario hacer más trabajo en la conceptuación y definición cuidadosa de los constructos que puedan ubicarse en un marco teórico significativo y coherente.

LOCUS DE CONTROL

El constructo descrito como "locus de control" se hizo famoso con la publicación de una monografía de Rotter (1966), en la que presentó la escala que había desarrollado para evaluar las expectativas generalizadas del individuo sobre el control interno o externo del reforzamiento (Escala I-E). El instrumento se construyó en el contexto de la teoría del aprendizaje social. Al explicar, su uso Rotter escribió: "El efecto del reforzamiento que sigue a alguna conducta [...] no consiste sencillamente en un proceso de estampado, sino que depende de si el individuo percibe o no una relación causal entre su conducta y la recompensa" (1966, p. 1). El control interno se refiere a la percepción de que los acontecimientos dependen de una conducta o alguna característica relativamente permanente del individuo; por otro lado, el control externo indica que el individuo no percibe el reforzamiento positivo o el castigo que siguen a alguna acción como enteramente contingentes a su comportamiento, sino como el resultado del azar, el destino o la suerte; puede percibirse como si estuviera bajo el control de otras personas poderosas e impredecibles por la complejidad de las fuerzas que le rodean.

La Escala I-E es un inventario autodescriptivo de elección forzada, del que se presentan dos reactivos ilustrativos en la tabla 14.1. El lector puede encontrar en la monografía de Rotter (1966) la lista completa de reactivos junto con las instrucciones para su aplicación. La publicación original presenta mucha información concerniente a la escala I-E, incluyendo las normas percentilares de varios centenares de estudiantes de ambos géneros de una sola universidad, junto con las medias y *DE* de algunas docenas de otras muestras compuestas sobre todo por universitarios. Subsecuentemente, otros proyectos independientes de investigación han acumulado datos sobre muchos otros grupos, y también se dispone de un cuerpo considerable de datos sobre la validez de constructo. Los análisis factoriales indicaban al principio que un único factor general podía dar cuenta de la mayor parte de la varianza de la respuesta; sin embargo, más adelante otros análisis factoriales indicaron que el constructo podría subdividirse en varios factores identificables, como lo ilustra la creencia en un mundo difícil, un mundo injusto, un mundo impredecible y un mundo políticamente insensible (B. E. Collins, 1974). En muchos otros estudios posteriores se descubrió también estructuras multifactoriales.

Para mediados de la década de los setenta era claro que las expectativas de control del género que cubre la Escala I-E pueden ayudar a predecir ciertas conductas; sin embargo, los investigadores del campo también reconocen que para elevar al máximo la

Tabla 14.1

Dos reactivos ilustrativos de la escala Interno-Externo (I-E)

(a)	A la larga, la gente obtiene en este mundo el respeto que merece.
(b)	Desafortunadamente, el esfuerzo de uno a menudo queda sin reconocimiento, sin importar lo mucho que se esfuerce.
(a)	Tener éxito es cuestión de trabajo duro; la suerte tiene poco o nada que ver con ello.
(b)	Obtener un buen trabajo depende principalmente de estar en el lugar correcto en el momento indicado.

Las instrucciones afirman: "Éste es un cuestionario para averiguar la forma en que algunos acontecimientos importantes de nuestra sociedad afectan a personas diferentes [...] Seleccione por favor la afirmación de cada par (y *sólo una*) que usted *crea* firmemente que es cierta en lo que a usted concierne."

(Tomado de Rotter, 1966, p. 11. Reproducido con autorización.)

precisión predictiva es necesario que las medidas de las expectativas de control se dirijan a las poblaciones específicas y las áreas de comportamiento estudiadas. Desde entonces, se han elaborado diversas escalas de locus de control, algunas de ellas adaptadas para usarse con diversas poblaciones, incluyendo niños de edad escolar y preescolar (véase, por ejemplo, Connell, 1985; Herzberger, Linney, Seidman y Rappaport, 1979; Nowicki y Duke, 1983); otras pretenden evaluar el control de creencias respecto a campos específicos, como la satisfacción marital o el estado de salud mental (D. J. Hill y Bale, 1980; P. C. Miller, Lefcourt y Ware, 1983); otras más abarcan las creencias causales sobre diferentes áreas (por ejemplo, logro y afiliación) o diferentes esferas de control, como la eficacia personal, el control interpersonal y el control sociopolítico (Lefcourt, von Baeyer, Ware y Cox, 1979; Paulhus, 1983). El lector puede encontrar en Lefcourt (1991) la reproducción de varias escalas con los datos psicométricos básicos.

El número y la gama de instrumentos disponibles para evaluar el locus de control atestiguan la vitalidad del constructo. Entre 1984 y 1995 se encontraron más de 5 000 registros con este tema en la base de datos PsycINFO. Las investigaciones sobre la función de las expectativas de control en las prácticas relacionadas con la salud son especialmente numerosas. Además, el locus de control es un aspecto importante de la motivación en general y está muy relacionado con otras áreas importantes de la investigación de la personalidad, como la teoría de la atribución, el desamparo aprendido y la autoeficacia, entre muchos otros (Skinner, 1995).

Técnicas proyectivas

La cantidad de técnicas proyectivas es grande y está diversificada. En este capítulo, consideramos algunos de los ejemplos más conocidos, pero, a excepción de ciertos puntos que son exclusivos de algunas técnicas, no haremos una evaluación crítica de ningún instrumento en particular, sino que, en una sección aparte, haremos una apreciación resumida de las técnicas proyectivas en la que habremos de destacar los problemas metodológicos comunes. Las técnicas proyectivas representan una curiosa discrepancia entre la investigación y la práctica, ya que aunque suelen desempeñar un mal papel en las evaluaciones, su popularidad en la práctica clínica no disminuye (Bellak, 1992; Lubin, Larsen y Matarazzo, 1984; Piotrowski, 1984; Piotrowski, Sherry y Keller, 1985; Piotrowski y Zalewski, 1993; Watkins, 1991). En la última sección, examinaremos la naturaleza y las implicaciones de esta inconsistencia.

La bibliografía relativa a las técnicas proyectivas es vasta y llega a abarcar más de 6 000 referencias de un solo instrumento. El lector interesado en las técnicas proyectivas disponibles puede consultar las obras de Klopfer y Taulbee (1976), Rabin (1981, 1986) y Reynolds y Kamphaus (1990b, capítulos 3 a 8). Los Anuarios de Medición Mental (*Mental Measurements Yearbooks*, MMY) contienen revisiones de la mayor parte de los instrumentos actuales, y aunque el *Journal of Personality Assessment*[1] está abierto a artículos sobre todo tipo de pruebas, publica una gran cantidad de material sobre diversos aspectos del uso de las técnicas proyectivas.

[1] La publicación ha pasado por varios cambios de nombre desde que fue iniciada por Bruno Klopfer en 1936 como el boletín *Rorschach Research Exchange*; de 1950 a 1963 se conoció como el *Journal of Projective Techniques*.

NATURALEZA DE LAS TÉCNICAS PROYECTIVAS

Una característica importante de las técnicas proyectivas es el empleo que hacen de una tarea *no estructurada* en términos relativos, es decir, una tarea que permite una variedad prácticamente ilimitada de respuestas posibles. Para permitir el libre juego de la fantasía del individuo se proporcionan breves instrucciones generales, y, por la misma razón, los estímulos de la prueba suelen ser vagos o ambiguos. La hipótesis de fondo es que la forma en que el individuo percibe e interpreta el material del instrumento o en que "estructura" la situación reflejará aspectos fundamentales de su funcionamiento psicológico. En otras palabras, se espera que los materiales de la prueba habrán de servir como una especie de pantalla sobre la que los individuos "proyectan" sus procesos de pensamiento, necesidades, ansiedades y conflictos característicos.

Además, en la medida en que los examinados rara vez se percatan de la interpretación psicológica que se hará de sus respuestas, los instrumentos proyectivos suelen ser procedimientos *encubiertos* de evaluación. De igual modo, las técnicas proyectivas se caracterizan por aproximarse de manera *global* a la valoración de la personalidad, es decir, no concentran la atención en la medición de rasgos separados sino en una imagen compuesta de la personalidad. Por fin, los exponentes de las técnicas proyectivas en general las consideran especialmente eficaces para revelar los aspectos *encubiertos*, *latentes* o *inconscientes* de la personalidad; más aún, argumentan que entre menos estructurada esté la prueba, más sensible resulta al material encubierto, como se sigue de la suposición de que entre menos estructurados o más ambiguos sean los estímulos, menor será la probabilidad de que provoquen reacciones defensivas de parte del examinado.

Los métodos proyectivos se originaron en los medios clínicos, en los que continúan siendo una herramienta importante, y algunos evolucionaron de los procedimientos terapéuticos (como la terapia artística) empleados con los pacientes psiquiátricos. Aunque la mayor parte de las técnicas proyectivas refleja en su marco teórico la influencia de los conceptos psicoanalíticos tradicionales y modernos, se han hecho algunos intentos dispersos por establecer una base para las técnicas proyectivas en la teoría estímulo-respuesta y en las teorías perceptuales de la personalidad (véase, por ejemplo, Lindzey, 1961/1977). Por supuesto, debe advertirse que *no* es preciso evaluar las técnicas a la luz de sus orígenes históricos o sus inclinaciones teóricas particulares, pues cada procedimiento puede demostrar ser útil en la práctica o gozar de validez empírica por razones distintas a las que se citaron primero para justificar su introducción.

TÉCNICAS DE LAS MANCHAS DE TINTA

Rorschach. Una de las técnicas proyectivas más populares es la de Rorschach que se basa en la utilización de manchas de tinta (Aronow y Reznikoff, 1983; Aronow, Reznikoff y Moreland, 1994; Erdberg y Exner, 1984; Exner, 1993). La estableció el psiquiatra suizo Hermann Rorschach (1921/1942), y su primera descripción data de 1921. Aunque los psicólogos habían usado antes series estandarizadas de manchas de tinta en sus estudios sobre la imaginación y otras funciones, Rorschach fue el primero en aplicarlas a la investigación diagnóstica del conjunto de la personalidad. En el desarrollo de esta técnica, Rorschach experimentó con un gran número de

manchas de tinta, que presentaba a diferentes grupos psiquiátricos. Como resultado de sus observaciones clínicas, las respuestas características que diferenciaban entre los diversos síndromes psiquiátricos se incorporaron gradualmente en un sistema de calificación cuyos procedimientos fueron luego perfeccionados por la evaluación complementaria de personas normales y otras con retardo mental, así como de artistas, intelectuales y otros grupos. De esta manera, la metodología de Rorschach fue la primera aplicación informal y relativamente subjetiva de las claves de criterio.

Debido a la inesperada muerte de Rorschach en 1922, sus colegas y sus estudiantes continuaron con el desarrollo del test. En las décadas siguientes, el uso de la técnica de Rorschach se difundió tanto en Europa como en los Estados Unidos; sin embargo, en ausencia de un sistema único, los procedimientos de aplicación, calificación e interpretación de "*el* Rorschach" proliferaron y evolucionaron en varios métodos o sistemas.[2] Para la década de los sesenta, hablar del Rorschach como el único test estandarizado era, de hecho, inexacto, ya que los diversos sistemas y usuarios sólo compartían las 10 tarjetas de estímulo originales y algunos postulados interpretativos básicos derivados del trabajo original de Rorschach.

Cada una de las tarjetas del Rorschach tiene impresa una mancha de tinta con simetría bilateral similar a la ilustrada en la figura 15.1. Cinco de las manchas se presentan en tonos de gris y negro, dos contienen toques adicionales de un rojo brillante y las tres restantes combinan varios tonos al pastel. Durante la aplicación del instrumento se presentan las manchas, una tras otra, y se pide al individuo que diga lo que piensa que representan. Además de llevar un registro textual de las respuestas a cada tarjeta, los examinadores suelen anotar el tiempo de reacción y la duración de las respuestas, la posición o posiciones en que se sostienen las tarjetas, los comentarios espontáneos, las expresiones emocionales y otros comportamientos incidentales del individuo durante la sesión de prueba. En algún punto después de la presentación de las 10 tarjetas, casi todos los evaluadores cuestionan sistemáticamente al individuo respecto a partes y aspectos de la mancha con los que hizo las asociaciones, lo que le permite elaborar y aclarar sus primeras respuestas.

Los diversos sistemas Rorschach, que florecieron entre las décadas de los treinta y los sesenta, diferían principalmente en los métodos de calificación y, por ende, en materia de interpretación. En esencia, en la interpretación del Rorschach el foco de interés se concentra en el contenido de las respuestas o en sus características formales como la localización, los determinantes, la calidad de la forma y los diversos resúmenes cuantitativos derivados de las respuestas. Aunque los sistemas del Rorschach difirían en los detalles de calificación e interpretación de las respuestas a las tarjetas, muchos compartían la clasificación básica de las categorías de calificación. La *localización* se refiere a la parte de la mancha con la que el individuo asocia cada respuesta: ¿utiliza toda la mancha, un detalle común, un detalle inusual, un espacio en blanco o alguna combinación de las áreas? Los *determinantes* de la respuesta incluyen forma, color, sombreado y "movimiento", aunque desde luego en la mancha no hay movimiento, se califica en esta categoría la percepción que el individuo tiene de ella como la

[2] Véase Exner (1969, 1993) para una historia más detallada del Rorschach y su evolución.

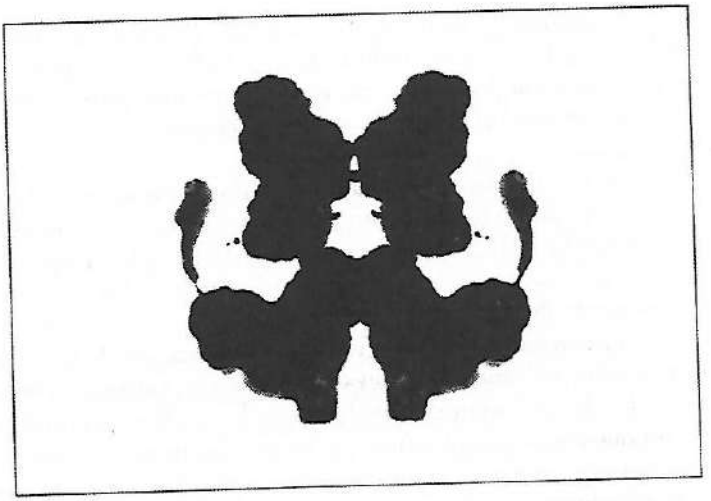


Figura 15.1. Mancha de tinta del tipo utilizado en la Técnica de Rorschach.

representación de un objeto que se mueve. En estas categorías se hacen diferenciaciones posteriores; por ejemplo, el movimiento humano, el movimiento animal y el movimiento abstracto o inanimado se califican por separado. De igual modo, el sombreado puede percibirse como si representara profundidad, textura, formas imprecisas como nubes o reproducciones acromáticas de colores como en las fotografías. La *calidad* o *nivel de la forma* de las respuestas puede referirse a la localización usada en la mancha, a su originalidad o a ambas. Además, en algunos sistemas también puede calificarse la complejidad cognoscitiva de las respuestas y otros aspectos cualitativos de las percepciones.

El tratamiento del *contenido* también ha variado entre los distintos sistemas Rorschach, aunque regularmente emplean ciertas categorías principales, entre las que destacan las figuras y los detalles humanos (o partes de las figuras humanas), y las figuras y detalles de animales. Otras categorías incluyen objetos de arte, plantas, mapas, nubes, sangre, rayos X, ropa, atributos sexuales y paisajes. A menudo se calcula una calificación popular de acuerdo con la frecuencia relativa de diferentes respuestas entre la gente en general. Para cada una de las 10 tarjetas, ciertas respuestas específicas se califican como populares por lo común de su ocurrencia; además, casi todos los sistemas incluyen un conteo de las verbalizaciones inusuales que hace el examinado durante el Rorschach, las cuales resultan particularmente útiles en la detección de formas graves de psicopatologías.

En general, el análisis de las respuestas al Rorschach se basa en la frecuencia relativa de éstas en las diversas categorías así como en ciertas razones y relaciones entre distintas categorías. Algunos ejemplos del tipo de interpretaciones cualitativas que suelen emplearse con las respuestas del Rorschach son la asociación de respuestas "globales" con el pensamiento conceptual, respuestas de "color" con la emocionalidad y de "movimiento humano" con la imaginación y la fantasía. En la aplicación del Rorschach, suele hacerse hincapié en la descripción "global" final del individuo, en la que

el clínico integra los resultados de diferentes partes del protocolo y toma en consideración las interrelaciones de diferentes puntuaciones e índices. En la práctica, también se usa la información derivada de fuentes externas (como otras pruebas, entrevistas y los registros de historias de caso) al preparar esas descripciones.

Sistema comprensivo de Exner Para la década de los sesenta el Rorschach había caído en descrédito como instrumento psicométrico. Los investigadores se encontraban en un callejón sin salida debido tanto a las dificultades inherentes al método (variabilidad en el número total de respuestas, influencia de los efectos del examinador e interdependencia de las calificaciones) como a la proliferación de sistemas de calificación. Estas circunstancias hicieron que la investigación de la confiabilidad y la validez del Rorschach resultara una empresa fragmentada, cargada de fallas metodológicas y de resultados decepcionantes. Muchos psicólogos clínicos continuaron usando el Rorschach, aunque casi todos admitían que no seguían un solo sistema y que tendían a usar los datos del instrumento a su manera, que podía fluctuar entre una interpretación cualitativa totalmente impresionista a la adherencia a uno o más sistemas, más o menos estrictos, según consideraran pertinente.

Las grandes diferencias entre los cinco principales sistemas Rorschach en uso en los Estados Unidos fueron documentadas por John E. Exner, Jr. (1969), quien trabajó con Samuel Beck y Bruno Klopfer, dos de los más importantes sistematizadores del Rorschach.[3] Luego de sus investigaciones tanto de la bibliografía como del uso clínico del Rorschach, Exner quedó fascinado por la posibilidad de fundir en un solo sistema todos los rasgos empíricamente defendibles y útiles que el método pudiera tener, por lo que durante los últimos 25 años ha conducido el más ambicioso y fructífero esfuerzo jamás emprendido para dar al Rorschach una base psicométrica adecuada (Exner, 1974, 1991, 1993, 1995; Exner y Weiner, 1995).

Exner desarrolló un sistema Rorschach que integraba elementos seleccionados de los cinco planteamientos principales y en el que ofrecía procedimientos estandarizados de aplicación, calificación e interpretación elegidos en función de las comparaciones empíricas entre diversas prácticas. El sistema subraya las variables estructurales más que las de contenido ya que, según Exner, la calificación de las respuestas tiene como propósito llegar a un *resumen estructural* que, además de ser el núcleo del sistema, establece la base para la mayor parte de los postulados interpretativos. Cada respuesta se codifica en diferentes categorías de calificación que incluyen localización, determinantes, calidad de la forma, contenidos, actividad organizacional y respuestas populares, entre otras. Luego de hacer una lista con las respuestas codificadas se obtiene la frecuencia de los códigos para luego utilizar estos elementos en el cálculo de razones, porcentajes e índices que completan el resumen estructural. La interpretación de las declaraciones puede derivarse, considerando todo el registro del Rorschach, para variables con diversos niveles de complejidad. Algunas hipótesis se vinculan con las simples frecuencias, como el grado en que se utiliza un solo determinante (por ejemplo, el sombreado); otras, se basan en la ocurrencia combinada de dos o más variables

[3] Los otros tres sistemas importantes son los de Marguerite Hertz, Zygmunt Piotrowski y David Rapaport con Roy Schafer.

(como el número de contenidos humanos y animales). El nivel más complejo de análisis incluye constelaciones de diversas variables y calificaciones de corte derivadas en forma empírica. Estas variables se agrupan en índices (por ejemplo, índice de esquizofrenia, índice de depresión e índice de déficit de afrontamiento) que supuestamente reflejan la probabilidad de que ciertos trastornos o condiciones estén presentes.[4]

Mediante el uso de este sistema uniforme, que ha evolucionado y se ha perfeccionado en las dos últimas décadas, Exner y sus colaboradores han recopilado una gran cantidad de datos psicométricos, que incluyen normas obtenidas en adultos, niños y adolescentes, así como en diversas muestras psiquiátricas de referencia. Los estudios de la confiabilidad en el retest realizados con intervalos que varían de unos cuantos días a tres años indican una considerable estabilidad temporal para la mayor parte de las variables calificadas. La cuidadosa descripción de las directrices de calificación del sistema comprensivo ha hecho posible que los examinadores capacitados obtengan altos porcentajes de acuerdo entre calificadores. De hecho, una de las principales contribuciones del trabajo de Exner es que logró un sistema Rorschach uniforme que permite la comparación de los descubrimientos empíricos de diferentes investigadores,[5] de ahí que no resulte sorprendente que su sistema se haya convertido en la aproximación más enseñada para calificar e interpretar el Rorschach y haya demostrado su utilidad al aumentar el poder estadístico de la investigación con dicho instrumento (véase, por ejemplo, Acklin, McDowell y Orndoff, 1992; Ritzler y Alter, 1986).

A pesar de las claras mejoras metodológicas que el sistema de Exner ha producido en el Rorschach, aún quedan por resolver varias interrogantes, de las cuales el tema de la validez es la principal y más complicada. La bibliografía sobre el tema no sólo es enorme, sino que también está llena de descubrimientos contradictorios. El Rorschach, igual que el MMPI, se ha utilizado para muy diversos propósitos, muchos de los cuales trascienden el objetivo inicial de su autor, lo que ha contribuido a aumentar la complejidad del estudio de la validez. En general, la investigación meta-analítica ha demostrado que los índices de validez convergente para el Rorschach son comparables a los que se consiguen con el MMPI (Atkinson, Quarrington, Alp y Cyr, 1986; K. C. H. Parker, Hanson y Hunsley, 1988). Además, el trabajo del propio Exner con el Rorschach da un considerable apoyo a la validez de muchos constructos evaluados con su sistema y a su utilidad para describir ciertos aspectos del funcionamiento de la personalidad. Sin embargo, cuando se trata de diagnosticar condiciones actuales complejas o predecir el comportamiento futuro, los resultados de la investigación suelen ser mixtos (Exner, 1996; Weiner, 1994a; Wood, Nezworski y Stejskal, 1996a, 1996b).

Una complicación importante en la interpretación de las calificaciones del Rorschach es el número total de respuestas, lo que se conoce como productividad de respuesta o "R". Cuando ocurren grandes diferencias en R entre individuos o grupos, es probable que sigan diferencias en otras categorías de calificación, por lo que las discre-

[4] En el sistema Comprensivo se tiene acceso a programas de calificación por computadora y de ayuda para la interpretación, que sólo requieren la codificación inicial de las respuestas.

[5] Sin embargo, aún no ha quedado establecido el tema de cómo cuantificar mejor el acuerdo entre calificadores en el Rorschach. Para un análisis del problema y la comparación de tres diferentes métodos de estimación de la confiabilidad de los datos del Rorschach, véase McDowell y Acklin (1996).

pancias encontradas en ciertas categorías pueden ser un artefacto que resulta de las variaciones en R. A esta característica intrínseca de las calificaciones del Rorschach debe agregarse el hecho de que la productividad de respuesta parece estar relacionada con otras variables, como el nivel intelectual y la escolaridad.[6] De manera característica, los investigadores del Rorschach difieren en sus opiniones de la importancia y grado de los problemas presentados por las variaciones en la productividad de respuesta; por ejemplo, G. J. Meyer (1992, 1993) ha señalado la necesidad de estudiar el significado psicológico de R y de comparar sus méritos contra las desventajas de controlar o ajustar esta variable; otros, consideran que ya se ha demostrado que su impacto es insignificante en la mayor parte de las circunstancias (Exner, 1992; Weiner, 1995b), mientras que algunos más sostienen que el problema planteado por R debe tratarse de diferente manera según el uso que se quiera hacer de los datos y, en el caso de los datos de investigación, dependiendo de qué tanto se aparte de la distribución de la normalidad (Kinder, 1992; Lipgar, 1992).

El sistema comprensivo de Exner no está exento de críticas. Muchos usuarios del Rorschach han objetado la naturaleza ateórica de su aproximación y el uso excesivo de los datos de contenido que, según argumentan, reduce significativamente su utilidad clínica. Los críticos también han advertido que el sistema es demasiado complejo y en ocasiones vago y contradictorio. Más aún, la investigación del mismo Exner presenta los defectos del reducido tamaño de sus muestras, el gran número de variables, la falta de estudios de validación cruzada y la inaccesibilidad al escrutinio público.[7] Sea como fuere, lo cierto es que la disponibilidad del sistema de Exner, junto con los datos de investigación que con sus colaboradores ha compilado, ha revitalizado el uso psicométrico del Rorschach.

Aproximaciones alternativas. A pesar de la gran aceptación del sistema comprensivo de Exner, coexisten otras aproximaciones; de hecho, muchos conceden al trabajo de Exner el crédito de haber renovado el interés en el Rorschach desde otros ángulos además del suyo. Aronow y sus colaboradores (Aronow y Reznikoff, 1976, 1983; Aronow *et al.*, 1994, 1995) describen una alternativa de orientación más clínica que en esencia considera al Rorschach una entrevista clínica estandarizada en la que se muestrean las operaciones perceptuales del sujeto. El sistema se concentra en la interpretación del contexto más que en las variables estructurales o los determinantes perceptuales de las respuestas, pero no considera que las escalas de contenido disponibles ni los sistemas de calificación sean instrumentos psicométricos lo bastante confiables para usarlos en el diagnóstico individual. Los autores recomiendan una aplicación estrictamente clínica del Rorschach como forma de mejorar la comprensión ideográfica del caso individual, y advierten que casi todos los clínicos experimentados

[6] Un descubrimiento interesante, relacionado con R, de la investigación sobre la susceptibilidad del Rorschach a la simulación es que, cuando se pide a los sujetos que finjan psicosis, hay una reducción en la productividad de la respuesta (G.G. Perry y Kinder, 1990).

[7] Para una revisión de esas y otras críticas al trabajo de Exner, véase Aronow, Reznikoff y Moreland (1995), Kleiger (1992), P. M. Lerner (1994), W. Perry, (1993), Viglione (1989), Vincent y Harman (1991), y Wood, Nezworski y Stejskal (1996a, 1996b).

orientan sus pasos hacia esta aproximación por su utilidad en el proceso de psicoterapia. Su interpretación se basa ante todo en el contenido de las respuestas, que complementan con el comportamiento verbal y no verbal. De acuerdo con la investigación disponible y con la experiencia clínica, Aronow y sus colaboradores prepararon un conjunto de directrices para obtener interpretaciones ideográficas más eficaces y confiables; por ejemplo, sugieren que en el caso individual es más probable que sean significativas las respuestas que se apartan del lugar común y las que están más vinculadas a las propiedades de estímulo de ciertas manchas. De igual modo, previenen en contra de los sistemas rígidos de interpretación simbólica que atribuyen significados fijos a las categorías de contenido o asignan cualidades evocativas invariables a las manchas del Rorschach. En lugar de ello, ofrecen procedimientos para investigar el significado de las respuestas de una manera conservadora que es compatible con los principios generales del psicodiagnóstico y que hace uso de las experiencias del individuo.

Paul Lerner (1991) propone otra aproximación al Rorschach que, en contraste con la posición ateórica de Exner, está impregnada de la moderna teoría psicoanalítica de la que evolucionó desde la década de los setenta. Mientras que Exner (1989) afirma que el Rorschach es un instrumento en el que la proyección rara vez participa, Lerner considera que se trata de un método fundamentalmente proyectivo que brinda acceso al mundo interno del individuo.[8] El texto de Lerner proporciona una guía para los usos clínicos del Rorschach y las aplicaciones de su investigación a la evaluación de las representaciones del objeto, las maniobras defensivas y otros conceptos centrales de las teorías psicodinámicas modernas.

Una aplicación clínica especial del Rorschach es la que ilustra el consenso Rorschach (Aronow *et al.*, 1994, capítulo 13; Blanchard, 1968; Cutter y Farberow, 1970). En esta adaptación, las manchas de tinta se presentan para la interpretación conjunta de parejas u otros miembros de la familia, colaboradores, miembros de pandillas juveniles u otros grupos naturales. Mediante la discusión y la negociación, los participantes deben llegar a un acuerdo sobre un solo conjunto común de respuestas. La técnica ha sido empleada con aparente éxito como base para explorar las relaciones interpersonales y otras clases de comportamiento social.

El Rorschach se ha descrito como "una prueba que ha sobrevivido repetidamente a sus obituarios" (Peterson, 1994, p. 396). En muchas ocasiones se ha anticipado su desaparición porque, al igual que otras pruebas psicológicas de gran uso, ha sido muy mal empleada. En la actualidad, la técnica experimenta un notable resurgimiento tanto en términos de la actividad de investigación como en la teorización. Aunque persisten diferentes formas de aproximarse al Rorschach, los usuarios parecen convencidos de que tiene un valor especial para estudiar los aspectos perceptuales, cognoscitivos y afectivos del funcionamiento de la personalidad. Algunos prefieren considerarlo un método para generar datos que pueden ser usados desde diversos puntos de vista, mientras que otros se esfuerzan por vincular las tradiciones empíricas y teóricas así como diferentes aproximaciones teóricas al Rorschach en un sistema completamente integrado (véa-

[8] Las razones de esta aparente contradicción tienen sus raíces en los puntos de vista divergentes que dichos autores tienen respecto a la naturaleza de la proyección y de la tarea del examinado en el Rorschach.

se, por ejemplo, Acklin, 1995; Blatt, 1990; P. M. Lerner, 1994; Meloy y Singer, 1991; Weiner, 1994b; Willock, 1992).

La técnica de manchas de tinta de Holtzman. Incluso antes de que Exner comenzara a trabajar con su sistema comprensivo, Wayne H. Holtzman hizo un serio esfuerzo por aplicar la orientación psicométrica a la técnica de las manchas de tinta, tomando al Rorschach como modelo. La técnica de manchas de tinta de Holtzman (*Holtzman Inkblot Technique*, HIT) se diseñó para eliminar las principales deficiencias técnicas de los primeros instrumentos (Holtzman, 1961, 1986; Holtzman, Thorpe, Swartz y Herron, 1961); sin embargo, los cambios en los materiales de estímulo y los procedimientos son lo bastante grandes para que el instrumento de Holtzman se considere una prueba diferente y se evalúe sin referencia al Rorschach. La técnica de Holtzman proporciona dos series paralelas de 45 tarjetas cada una. Las manchas se seleccionaron de un banco preliminar en función de criterios empíricos que pretendían aumentar al máximo su efectividad. Se obtiene una sola respuesta por tarjeta. Hay tarjetas acromáticas y coloreadas, y algunas de las manchas son claramente asimétricas.

La aplicación y calificación de la HIT están bien estandarizadas y se describen con claridad desde el inicio; además, se obtienen calificaciones en 22 variables de respuesta, que incluyen muchas que igualan a las del Rorschach y algunas variables adicionales, como ansiedad y hostilidad. Para cada variable, se dispone de puntuaciones percentilares para muestras normales de niños y adultos, pero también para algunos grupos desviados (E. F. Hill, 1972; Holtzman, 1975). La confiabilidad entre calificadores parece ser muy satisfactoria; los estudios de la confiabilidad de división por mitades, formas alternas y test-retest han mostrado diferencias entre las variables de respuesta, aunque la mayor parte de los resultados son alentadores. Una forma grupal de la prueba, que utiliza transparencias, arroja calificaciones en casi todas las variables que son comparables a las obtenidas mediante la aplicación individual (Holtzman, Moseley, Reinehr y Abbott, 1963; Swartz y Holtzman, 1963). Holtzman (1988) propuso recientemente la HIT 25, una versión corta que se encuentra en proceso de establecimiento de normas (Swartz, 1992) y que está compuesta por las 25 primeras tarjetas de la forma A de la HIT con dos respuestas por tarjeta.

En la actualidad se cuenta con una cantidad considerable de datos sobre la validez de la HIT, casi todos alentadores (Gamble, 1972; Holtzman, 1975, 1986, 1988; Leichsenring, 1991; Sacchi y Richaud de Minzi, 1989; Swartz, 1973). En la investigación de la validación se han seguido diversos métodos, incluidos el estudio de las tendencias de desarrollo, comparaciones transculturales, correlaciones con otros instrumentos y con indicadores conductuales de las características de personalidad, así como comparaciones de grupos contrastados de individuos normales y pacientes psiquiátricos. Hay un manual preparado por E. F. Hill (1972) que se dirige en particular al uso clínico de la HIT.

Es claro que la HIT tiene algunas ventajas psicométricas sobre el Rorschach, pues la disponibilidad de formas paralelas no sólo permite medir la estabilidad en el retest, sino que también es adecuada para los estudios de seguimiento. La restricción a una respuesta por tarjeta mantiene constante para cada individuo la productividad de respuesta (R), con lo que se evitan muchos de los errores de calificación del Rorschach;

sin embargo, debe advertirse que todavía no se logra controlar la longitud de las respuestas (número de palabras) que, como en el Rorschach, ha demostrado relacionarse significativamente con varias calificaciones de la HIT (Megargee, 1966). A pesar de sus ventajas, la información sobre la HIT es escasa en comparación con la del Rorschach, por lo que se necesitan más datos para establecer la importancia diagnóstica de varias calificaciones y la validez de constructo de las variables de personalidad evaluadas con esta técnica (para revisiones, véanse Cundick, 1985; Dush, 1985).

TÉCNICAS PICTÓRICAS O GRÁFICAS

Test de Apercepción Temática. En contraste con las técnicas de manchas de tinta, el Test de Apercepción Temática (TAT) presenta estímulos muy estructurados y requiere respuestas verbales más complejas y organizadas en forma significativa. La interpretación que el examinador hace de las respuestas por lo general se basa en el análisis de contenido de una naturaleza más bien cualitativa. Desarrollado inicialmente por Henry Murray y su equipo en la Clínica Psicológica de Harvard (Murray *et al.*, 1938), el TAT no sólo se ha empleado de modo extenso en la investigación y la práctica clínica, sino que también ha servido como modelo para el desarrollo de muchos otros instrumentos (J. W. Atkinson, 1958; Bellak, 1993; Dana, 1996b; R. Harrison, 1965; Holmstrom, Silber y Karp, 1990; Klopfer y Taulbee, 1976, pp. 554–558; Obrzut y Boliek, 1986).

Los materiales del TAT consisten en 19 tarjetas que contienen dibujos no definidos en blanco y negro y una tarjeta en blanco.[9] Al individuo se le pide que elabore una historia que se ajuste al dibujo, que diga qué condiciones culminaron en el acontecimiento mostrado en el dibujo, que describa lo que sucede en el momento y lo que los personajes piensan y sienten y que dé el resultado. En el caso de la tarjeta en blanco, se le dan instrucciones de que imagine algún dibujo, que lo describa y que cuente una historia al respecto. El procedimiento original descrito por Murray en el manual requiere dos sesiones de una hora, en cada una de las cuales se emplean 10 tarjetas. Las tarjetas reservadas para la segunda sesión se eligieron deliberadamente de modo que fueran más inusuales, dramáticas y extrañas, y las instrucciones correspondientes exhortan a los individuos a que den rienda suelta a su imaginación. Se dispone de cuatro conjuntos de 20 tarjetas, para niños, niñas, y para varones y mujeres mayores de 14 años. La mayoría de los clínicos utiliza conjuntos abreviados de tarjetas especialmente seleccionadas, por lo que rara vez presentan más de 10 tarjetas a un solo examinado.

En el método original para interpretar las historias del TAT (Murray *et al.*, 1943), el examinador primero determina quién es el "héroe", el personaje de cualquier género con el que supuestamente se identifica el examinado. Enseguida, se analiza el contenido de las historias, haciendo referencia sobre todo a la lista de "necesidades" y "presiones", determinadas Murray (en el capítulo 13 explicamos algunas de las necesidades propuestas en relación con el Inventario de Preferencias Personales de Edwards). Los ejemplos comprenden el logro, la afiliación y la agresión. Las *presiones* se refieren a las fuerzas ambientales que pueden facilitar o interferir con la satisfacción de las necesi-

[9] El lector puede encontrar en W. G. Morgan (1995) una interesante descripción histórica del origen del TAT.

dades. Ser atacado o criticado, recibir afecto, ser confortado y la exposición al peligro físico son ejemplos de presiones. Al evaluar la importancia para el individuo de la fuerza de una necesidad o presión, se presta especial atención a la intensidad, duración y frecuencia de su ocurrencia en las diferentes historias, así como la singularidad de su asociación con el dibujo presentado. Se supone que es más probable que para el individuo tengan significado los materiales inusuales que se alejan de las respuestas comunes a cada dibujo.

Se ha publicado una buena cantidad de información normativa respecto de las características de respuesta más frecuentes para cada tarjeta, incluyendo la forma en la que cada una se percibe, los temas desarrollados, los papeles asignados a los personajes, los tonos expresados, la velocidad de las respuestas, la longitud de las historias y cosas por el estilo (J. W. Atkinson, 1958; W. E. Henry, 1956; Murstein, 1972). Aunque esos datos normativos proporcionan un marco general de referencia para interpretar las respuestas individuales, la mayoría de los clínicos confía en "normas subjetivas" que ha desarrollado a lo largo de su propia experiencia con el instrumento y en el conocimiento que ha adquirido del examinado por otros medios. Se ha generado una serie de esquemas cuantitativos y de escalas de valoración que producen una buena confiabilidad del calificador, pero es raro que se utilicen en la práctica clínica porque su uso requiere mucho tiempo. Aunque generalmente se aplica como prueba individual oral en las situaciones clínicas, el TAT también puede aplicarse por escrito y como prueba colectiva.

El TAT se ha empleado mucho en la investigación de la personalidad, pero, desafortunadamente, la diversidad de los procedimientos de aplicación y calificación e incluso de los materiales de estímulo asociados se ha extendido a las formas de investigación y a la práctica clínica (Keiser y Prather, 1990), y tal diversidad ha hecho muy difícil investigar las propiedades psicométricas "del TAT" como prueba psicológica distinta porque el método carece del requisito básico de uniformidad. Además, se dispone de una cantidad considerable de datos experimentales que demuestran que condiciones como el hambre, la privación de sueño y la frustración social afectan significativamente las respuestas al TAT (J. W. Atkinson, 1958). Aunque brinda apoyo a la hipótesis proyectiva, la sensibilidad del TAT a esas condiciones temporales puede complicar el significado de las respuestas. La cuestión de la consistencia interna de las respuestas del TAT también ha recibido atención (J. W. Atkinson, 1981; Entwisle, 1972), y se tiene clara la necesidad de aplicar cierto control a la longitud de las historias, o productividad, un problema que el TAT comparte con el Rorschach (J. W. Atkinson y Raynor, 1974, capítulo 3).

No obstante, no se cuestiona la utilidad de las técnicas de apercepción temática en general y del TAT en particular. Investigaciones recientes confirman la utilidad clínica de las diversas versiones del TAT lo mismo en las aplicaciones tradicionales como la evaluación del grado de psicopatología y el uso de los mecanismos de defensa, que en los usos recientes como la evaluación de las habilidades para la resolución de problemas (Cramer y Blatt, 1990; Hibbard *et al.*, 1994; Ronan, Colavito y Hammontree, 1993; Ronan, Date y Weisbrod, 1995). Una de las aplicaciones más prometedoras del TAT se encuentra en las escalas recién desarrolladas para la evaluación clínica de las relaciones objetales (Alvarado, 1994; Barends, Westen, Leigh, Silbert y Byers, 1990; Freedenfeld, Ornduffy Kelsey, 1995; Westen, 1991; Westen, Lohr, Silk, Gold y Kerber,

1990). Por otro lado, la utilidad del TAT tampoco se limita al análisis temático de las respuestas, ya que las características formales de la estructura y el contenido de las historias también pueden emplearse de manera fructífera en el estudio de individuos y de grupos (véase, por ejemplo, Cramer, 1996; McGrew y Teglasi, 1990; Teglasi, 1993).

Adaptaciones del TAT y pruebas afines. Se han desarrollado muchas adaptaciones del TAT para propósitos especiales cuya semejanza con el instrumento original varía, por lo que resulta arbitraria la decisión de dónde trazar la línea entre las versiones modificadas del TAT y los nuevos instrumentos basados en la misma aproximación general. Existen varias versiones del TAT que pueden emplearse para medir las actitudes hacia los problemas laborales, los grupos minoritarios, la autoridad y cosas similares (D. T. Campbell, 1950; R. Harrison, 1965), mientras que otras adaptaciones se orientan a la consejería vocacional, la valoración ejecutiva o bien se emplean en diversos proyectos de investigación. También se han elaborado formas para poblaciones especiales, incluidos niños preescolares, de educación elemental, niños discapacitados, adolescentes y diversos grupos étnicos y nacionales (R. Harrison, 1965).

Algunas adaptaciones al TAT se han concentrado en la medición exhaustiva de una sola necesidad o impulso, como el sexo o la agresión. Resulta en especial interesante la extensa investigación sobre la necesidad de logro (n-l) conducida durante unos 30 años por McClelland, Atkinson y sus colaboradores (J. W. Atkinson, 1958; J. W. Atkinson y Feather, 1966; J. W. Atkinson y Raynor, 1974; McClelland, 1985; McClelland, Atkinson, Clark y Lowell, 1953/1976). Para medir esta necesidad, se emplearon cuatro dibujos, dos de los cuales fueron tomados del TAT, y se desarrollaron esquemas detallados para calificar las historias resultantes en relación con las expresiones de la necesidad de logro. La técnica ha sido empleada en un amplio programa de investigación sobre la motivación de logro. Los problemas investigados van de la teoría básica de la motivación (J. W. Atkinson y Feather, 1966) a los orígenes sociales y las consecuencias de la necesidad de logro y su función en el surgimiento y la caída de las sociedades (McClelland, 1961/1976). Los meta-análisis de los estudios que comparan al TAT con los cuestionarios que miden la necesidad de logro proponen que ambos métodos son válidos, aunque para propósitos diferentes y para distintos aspectos de la evaluación del impulso de logro (Spangler, 1992).

Charles Smith, en asociación con John W. Atkinson, David C. McClelland y Joseph Veroff (1992), preparó un compendio de los sistemas de calificación utilizados en el análisis de contenido del material verbal en el que se incluyen los sistemas de calificación con una larga tradición en la investigación (por ejemplo, los que se prepararon para los motivos de logro, afiliación y poder), y muchos otros que abordan temas tan diversos como la ideología política y la capacidad de afrontamiento. Se analizan los temas conceptuales así como las consideraciones metodológicas relacionadas con el muestreo, la calificación y el análisis del material verbal. Aunque muchos de los sistemas descritos se basan en las modificaciones del TAT y reflejan las opiniones de Murray, también están representados muchos otros acercamientos teóricos. Más aún, los sistemas incluidos se dirigen al análisis del contenido manifiesto más que al contenido simbólico de las muestras de pensamientos (en contraste con las "proyecciones") y más a la investigación que al uso clínico.

Aunque se afirma que la versión original del TAT puede aplicarse incluso a niños de cuatro años,[10] el Test de Apercepción Infantil (*Children's Apperception Test*, CAT) se diseñó especialmente para aplicarse a niños de entre tres y 10 años de edad (Bellak, 1993). Las tarjetas del CAT sustituyen las tarjetas con figuras humanas por otras de animales, por el supuesto de que a los niños pequeños les resulta más sencillo identificarse en imágenes de animales que de seres humanos. Los diversos animales de las tarjetas se presentan en situaciones humanas, a la manera antropomórfica de las tiras cómicas y de los libros infantiles. Los dibujos están diseñados para evocar fantasías relacionadas con los problemas de la alimentación y otras actividades orales, la rivalidad entre hermanos, las relaciones entre el niño y sus padres, la agresión, el entrenamiento de esfínteres y otras experiencias infantiles. Los autores del CAT prepararon una modificación humana de la prueba (CAT-H) destinada a niños mayores, en especial para aquellos con una edad mental mayor de 10 años (Bellak y Hurvich, 1966). Los autores sostienen que la forma humana o animal puede ser más eficaz, dependiendo de la edad y la personalidad característica del niño (para revisiones, véase Hatt, 1985; Shaffer, 1985).

Una prueba desarrollada más recientemente, el Test de Apercepción para Niños de Roberts (*Roberts Apperception Test for Children*, RATC) supera a otras técnicas proyectivas en lo que respecta a la satisfacción de los estándares psicométricos para la construcción de pruebas y la evaluación (McArthur y Roberts, 1982; véase también Sines, 1985). El RATC proporciona dos conjuntos de 16 tarjetas de estímulo, uno para niños y otro para niñas; también se dispone de un conjunto complementario con dibujos de niños negros, pero éste no ha sido normalizado. Los dibujos se eligieron para representar situaciones familiares interpersonales que incluyan a los pequeños en sus relaciones con adultos o con otros infantes (véase la figura 15.2), y las historias se califican sobre una serie de escalas que cubren los problemas por los que los menores suelen ser llevados a las clínicas. Los fundamentos claros y explícitos permiten una calificación bastante objetiva de las respuestas. Las normas se basan en las respuestas de 200 niños considerados por sus maestros como bien ajustados cuyas respuestas se compararon con las de 200 pequeños atendidos en las clínicas de orientación infantil para obtener algunos de los datos de validación citados en el manual. Es evidente que el instrumento representa un serio esfuerzo por combinar la flexibilidad de las técnicas proyectivas con los procedimientos estandarizados de aplicación, calificación y evaluación de la prueba. Las investigaciones sobre la validez del RATC para diversos usos continúa mostrando resultados favorables (véase, por ejemplo, Palomares, Crowley, Worchel, Olson y Rae, 1991). Además, Glen E. Roberts (1994) preparó un manual con fundamentos detallados para la calificación e interpretación del RATC en el uso clínico.

TEMAS son las siglas de *Tell-Me-A-Story* ("Cuéntame un cuento"), un instrumento diseñado especialmente para la evaluación de las características afectivas, cognoscitivas y de personalidad de niños de cinco años a jóvenes de 18 (Constantino, Malgady y Rogler, 1988). El TEMAS emplea dos conjuntos paralelos de estímulos en tarjetas coloreadas, uno para los niños pertenecientes a minorías étnicas y el otro para blancos. Los materiales de estímulo se prepararon cuidadosamente para facilitar la produc-

[10] Para información específica relativa al uso clínico del TAT y otras técnicas narrativas en niños y adolescentes, véase Teglasi (1993) y Worchel y Dupree (1990).

Figura 15.2. Uno de los dibujos empleados en el Test de Apercepción para Niños, de Roberts.

(Copyright © 1982 por Western Psychological Services. Reproducido con autorización.)

ción verbal y fomentar narraciones que abordaran elecciones entre metas en conflicto, como una gratificación inmediata frente a una exigencia. El conjunto asignado a las minorías presenta personajes cuyos rasgos y tez morena sugieren un origen hispano o negro. Aunque se ha elogiado al TEMAS por el avance que se supone que posee sobre las tarjetas originales del TAT en cuanto a su adecuación para niños afroamericanos e hispanoamericanos, las propiedades psicométricas del instrumento, sobre todo en la confiabilidad test-retest y la consistencia interna, han sido cuestionados repetidamente (para revisiones, véase Dana, 1993, capítulo 8; Lang, 1992; Ritzler, 1993a).

Hay otros instrumentos similares de apercepción temática para los ancianos, entre los que se incluyen el Test de Apercepción Gerontológica (*Gerontological Apperception Test*, Wolk y Wolk, 1971) y el Test de Apercepción para Ancianos (*Senior Apperception Test*, Bellak, 1993; Bellak y Bellak, 1973). Ambos instrumentos emplean conjuntos de tarjetas que presentan a una o más personas de edad e ilustran problemas que pueden ser de interés para los ancianos, como la soledad, las dificultades familiares y el desamparo. De ambos, se ha criticado su publicación prematura y el uso de imágenes que tienden a perpetuar los estereotipos adversos del envejecimiento (J. P. Schaie, 1978; K. W. Schaie, 1978), además no han demostrado tener ventajas sobre el TAT en la evaluación de personas mayores (Fitzgerald, Pasewark y Fleisher, 1974; Foote y Kahn, 1979). El Test de Apercepción Gerontológica está fuera de circulación.

El Estudio Pictórico de Frustración de Rosenzweig. El TAT y algunas otras de las técnicas afines que acabamos de considerar utilizan imágenes para estimular la

fantasía y provocar respuestas verbales elaboradas. En contraste, en el Estudio Pictórico de Frustración, de Rosenzweig, Estudio P-F (*Rosenzweig Picture-Frustration Study, P-F Study*), que describimos en esta sección, la cobertura está más circunscrita y requiere respuestas más sencillas. El instrumento se encuentra disponible en formas para adultos, individuos de 14 años y mayores (Rosenzweig, 1950, 1978a, 1978d), para adolescentes de 12 a 18 años (Rosenzweig, 1970, 1976b, 1981a), y para niños de cuatro a 13 años (Rosenzweig, 1960, 1977, 1981b, 1988). El Estudio P-F, que se derivó de la teoría del autor sobre la frustración y la agresión, presenta una serie de caricaturas en las que una persona frustra a otra o llama la atención sobre una condición frustrante. En la figura 15.3 se presentan dos de esas caricaturas, que se tomaron de la forma para niños. En el espacio en blanco el individuo escribe lo que la persona frustrada podría contestar.

En el Estudio P-F, las respuestas se clasifican de acuerdo con el tipo y la dirección de la agresión. Las formas de agresión incluyen: obstáculo-dominancia, énfasis en el objeto frustrante, defensa del yo, concentrar la atención en la protección de la persona frustrada, persistencia de la necesidad, concentrarse en la solución constructiva del problema frustrante. La dirección de la agresión se califica como exoagresiva o dirigida hacia el ambiente externo; endoagresiva, o dirigida hacia uno mismo, e imagoagresiva, o la desviación de la agresión en una tentativa por encubrir o evadir la situación. Al calificar la prueba, se compara el porcentaje de respuestas que cae en cada una de estas categorías con los porcentajes normativos correspondientes. También es posible obtener la calificación de conformidad con el grupo (CCG), que muestra la tendencia del individuo a dar respuestas que coincidan con las respuestas modales de la muestra de estandarización.

Figura 15.3. Reactivos característicos de la forma infantil del Estudio Pictórico de Frustración de Rosenzweig.

(Copyright © 1976, por Saul Rosenzweig. Reproducido con autorización.)

Al ser de una cobertura más limitada y más estructurada y contar con procedimientos de calificación relativamente objetivos, el Estudio P-F se presta más a un análisis estadístico que la mayor parte de las otras técnicas proyectivas. Desde su inicio se han hecho esfuerzos sistemáticos por obtener normas y supervisar su confiabilidad y validez. Después de casi 50 años, tanto el autor como otros investigadores han realizado una considerable cantidad de investigaciones con el Estudio P-F, cuya bibliografía aborda las propiedades psicométricas del instrumento y temas como el diagnóstico clínico, los cambios en el desarrollo, las diferencias de género, las diferencias culturales y la relación entre el humor y la agresión (Graybill, 1990, 1993; Nevo y Nevo, 1983; Rosenzweig, 1976a, 1978b, 1978c; Rosenzweig y Adelman, 1977; para revisiones, véase Viglione, 1985 y Wagner, 1985).

TÉCNICAS VERBALES

Aunque todos los instrumentos proyectivos revisados hasta ahora requieren respuestas verbales, ciertas técnicas proyectivas son completamente verbales pues sólo utilizan palabras tanto en los materiales de estímulo como en las respuestas. Algunas de estas técnicas verbales pueden emplearse ya sea de manera oral o por escrito, pero todas pueden aplicarse por escrito a grupos; estas condiciones impiden su uso con niños pequeños o con personas iletradas o que hablan otro idioma.

Una técnica que se anticipó por más de medio siglo al torrente de las pruebas proyectivas fue la *prueba de asociación de palabras*, conocida originalmente como "prueba de asociación libre". Originalmente descrita por Galton (1879), más tarde Wundt y J. McK. Cattell la llevaron al laboratorio psicológico donde la adaptaron a diversos usos. El procedimiento comprende sencillamente la presentación de una serie de palabras no relacionadas, ante cada una de las cuales se pide al individuo que responda con la primera palabra que le venga a la mente. Los primeros psicólogos experimentales, así como los primeros evaluadores mentales, vieron en estas pruebas de asociación una herramienta para la exploración de los procesos de pensamiento.

La aplicación clínica del método de asociación de palabras fue estimulada por el movimiento psicoanalítico, aunque otros psiquiatras, como Kraepelin las habían investigado antes. Entre los psicoanalistas, la contribución de Jung al desarrollo de la prueba de asociación de palabras es la más notable. Jung (1910) seleccionó palabras-estímulo que representaran "complejos emocionales" comunes y analizó las respuestas relacionándolas al tiempo de reacción, el contenido y las expresiones físicas de tensión emocional. Más de 30 años después, Rappaport y sus colaboradores (1946/1968) desarrollaron en la clínica Menninger una técnica similar que, a juicio de sus autores, tenía un propósito doble: ayudar a detectar deterioros en los procesos de pensamiento y proponer áreas significativas de conflicto. También cabe mencionar el uso de la técnica de asociación de palabras como un "detector de mentiras",[11] una aplicación iniciada también por Jung que a partir de entonces fue tema de una extensa investigación tanto en el laboratorio como en situaciones prácticas (Burtt, 1931; Lindsley,

[11] La técnica de asociación de palabras ya no se usa de esta manera. Para una revisión de las aplicaciones más recientes del "detector de mentiras" o método del polígrafo en los medios industriales y organizacionales, véase el capítulo 17.

1955). La lógica que justificaba la asociación de palabras en la detección de mentiras o de la culpa era similar a la aplicada cuando se empleaba para descubrir áreas de conflicto emocional.

El trabajo inicial de Kent y Rosanoff (1910) ilustra un acercamiento diferente a la prueba de asociación de palabras. Diseñado primero como instrumento de identificación psiquiátrica, la Prueba de Asociación Libre de Kent-Rosanoff utilizaba un sistema de calificación completamente objetivo. Las palabras de estímulo consistían en 100 palabras neutrales comunes elegidas porque tendían a provocar las mismas asociaciones en la gente; por ejemplo, a la palabra *mesa* casi toda la gente respondía "silla" y ante *oscuro* decía "luz". Prepararon un conjunto de tablas de frecuencias, una para cada palabra-estímulo, que mostraba el número de veces que se daba cada respuesta en una muestra de 1 000 adultos normales. Al calificar la prueba se derivó un "índice de comunalidad" a partir de las frecuencias de las respuestas de cada examinado. Las comparaciones de los psicóticos con los normales indicó que para los primeros el índice de comunalidad es menor.

Sin embargo, el uso diagnóstico de las técnicas de asociación libre disminuyó al resultar evidente que la frecuencia de las respuestas también varía con la edad, el nivel socioeconómico y educativo, los antecedentes culturales y regionales, la creatividad y otros factores, de ahí que la interpretación adecuada de los resultados requiera que se obtengan normas de muchos subgrupos y que se les actualice cada tanto, conforme evolucione el uso de las palabras. Además, la popularidad de los conceptos psicoanalíticos tradicionales que estimuló el desarrollo de esas técnicas también ha decrecido (Rabin y Zlotogorski, 1981); no obstante, la Kent-Rosanoff ha conservado su posición como herramienta común de laboratorio; se han obtenido normas adicionales en diversos países y ha sido empleada en la investigación de la conducta verbal y la personalidad (Goldfarb y Halpern, 1984; Isaacs y Chen, 1990; Jenkins y Russell, 1960; Palermo y Jenkins, 1963; Postman y Keppel, 1970; Van der Made-Van Bekkum, 1971).

El *completamiento de frases* es otra técnica proyectiva verbal de gran uso tanto en la investigación como en la práctica clínica (P. A. Goldberg, 1965; Haak, 1990: D. H. Hart, 1986; Lah, 1989). En términos de la longitud de las respuestas, la estructura y otros aspectos, las pruebas de completamiento de frases ocupan una posición intermedia entre las técnicas de asociación de palabras y las temáticas. Las palabras iniciales o el origen de la frase generalmente permiten una variedad prácticamente ilimitada de complementos posibles; los ejemplos incluyen: Mi ambición...; Las mujeres...; Lo que me preocupa...; Mi madre... El cuerpo de la oración se formula de modo que provoque respuestas relevantes del dominio de personalidad investigada. La flexibilidad de la técnica representa una de sus ventajas para propósitos clínicos y de investigación, aunque existen algunas formas estandarizadas para aplicaciones más generales.

Un ejemplo muy difundido es el Formulario de Frases Incompletas de Rotter (*Rotter Incomplete Sentences Blank*, RISB; Rotter y Rafferty, 1950), un instrumento que consta de 40 frases incompletas y da las siguientes instrucciones al examinado: "Complete las frases de modo que expresen *sus verdaderos sentimientos*. Intente hacerlo con cada una. Asegúrese de completar la frase". Cada frase completada se valora sobre una escala de siete puntos. Según el grado de adaptación o inadaptación que indique, la calificación puede hacerse más objetiva gracias a los ejemplos que ofrece el manual para cada puntuación. La suma de las valoraciones del individuo proporciona la puntuación de

adaptación total, que puede usarse con propósitos de identificación; además, el contenido de las respuestas puede examinarse clínicamente para obtener claves más específicas de diagnóstico. El manual revisado del RISB incluye información normativa actualizada y una revisión de las investigaciones realizadas desde 1950 (Rotter, Lah y Rafferty, 1992).

Se han elaborado muchas otras pruebas de completamiento de frases tanto para la evaluación de diferentes poblaciones objetivos como para diversos usos de diagnóstico y de investigación[12] (para descripciones de varios instrumentos tradicionales, véase D. H. Hart, 1986, Lah, 1989; Rabin y Zlotogorski, 1981). Algunas adiciones interesantes de reciente aparición abarcan instrumentos diseñados para detectar la simulación de la incapacidad durante los exámenes, para predecir la eficacia administrativa y para evaluar constructos, como los mecanismos de defensa, que pueden ser relevantes para la evaluación de la personalidad (Carson y Gilliard, 1993; N. L. Johnson y Gold, 1995; Timmons, Lanyon, Almer y Curran, 1993).

RECUERDOS AUTOBIOGRÁFICOS

Uno de los desarrollos recientes más prometedores en el área de las técnicas verbales proyectivas es el resurgimiento del interés en el uso de los *recuerdos autobiográficos* para la evaluación de la personalidad. Es obvio que el análisis de los recuerdos, en especial los más tempranos, para comprender conflictos recurrentes o intratables que se presentan en la vida posterior está ligado a la psicoterapia psicodinámica desde la época de Freud.[13] Además, Alfred Adler, que aparte de ser uno de los seguidores originales de Freud llegó a fundar su propia escuela de psicología individual, consideraba que la clave para entender el "estilo de vida" de la persona estaba en particular en el primer recuerdo, por lo que los psicólogos adlerianos utilizan desde los años treinta los recuerdos tempranos como herramientas clínicas y ocasionalmente en la investigación (véase, por ejemplo, Hafner, Fakouri y Labrentz, 1982; Slavik, 1991). Otros teóricos también han reconocido el papel central que en la evolución de la personalidad pueden desempeñar los recuerdos autobiográficos (considerados más como construcciones o proyecciones que como narraciones verdaderamente históricas). Pese a ello, luego de una corriente de interés en la primera parte del siglo XX, no fue sino hasta hace muy poco tiempo que esta fuente de información sobre la personalidad empezó a explorarse de manera constante y sistemática.

Como resultado de la influencia del punto de vista cognoscitivo, desde el inicio de los ochenta se observa un renovado interés en los recuerdos autobiográficos y, en particular, en la función que cumplen en la organización de la personalidad (Bruhn y Last, 1982; Ross, 1991; Rubin, 1986; Singer y Salovey, 1993). En el área destaca el trabajo de Arnold R. Bruhn (Bruhn, 1984, 1985, 1990a, 1990b), quien después de re-

[12] Un instrumento que ha sido ampliamente utilizado en la investigación es el Test de Completamiento de Oraciones de la Universidad de Washington (*Washington University Sentence Completion Test, WUSCT*). Diseñado para evaluar las etapas en el desarrollo del autoconcepto, lo estudiamos en el capítulo 16 junto con otras medidas del autoconcepto.

[13] El lector puede encontrar en Bruhn (1995a) una breve historia del uso de los recuerdos autobiográficos en la evaluación de la personalidad.

visar los modelos aplicados por los freudianos, los adlerianos y los psicólogos del yo a la interpretación de los primeros recuerdos, propuso un nuevo marco conceptual para éstos, así como una forma más sistemática de usarlos. En la teoría cognoscitivo-perceptual de Bruhn, los recuerdos autobiográficos son cruciales para comprender la personalidad, por lo que una de sus prioridades es el desarrollo de un método estandarizado para obtenerlos e interpretarlos. El Procedimiento de los Recuerdos Tempranos (PRT, Bruhn, 1989, 1992a, 1992b) es un instrumento autoaplicado de lápiz y papel que muestrea 21 recuerdos autobiográficos que no se restringen a la niñez, sino que se extienden a todo el ciclo vital. La primera parte requiere seis recuerdos generales o "espontáneos", delimitados sobre todo por los marcos temporales específicos (por ejemplo, los cinco recuerdos más tempranos y uno particularmente importante en el curso de la vida). La segunda parte comprende 15 recuerdos específicos o "dirigidos" que exploran un conjunto diverso de acontecimientos y áreas que pueden ser de relevancia clínica (digamos, un recuerdo traumático, el recuerdo del primer castigo recibido o uno de los recuerdos más felices).[14] Además de las descripciones narrativas de cada recuerdo, el PRT incluye diversas pruebas relativas a la claridad, el tono afectivo, la importancia y muchos otros elementos de los recuerdos. Bruhn considera los recuerdos tempranos de acontecimientos específicos como historias o metáforas que reflejan lo que la gente intuyó o aprendió conscientemente de sus experiencias; y aunque cree que esas historias a menudo son inexactas o están distorsionadas, sostiene que su veracidad carece de importancia para los propósitos clínicos porque, como en el caso de otras producciones proyectivas, el valor de los primeros recuerdos estriba en su capacidad para revelar preocupaciones, actitudes, creencias y estados afectivos actuales.

Aunque Bruhn y sus asociados desarrollaron, y revisaron, el Sistema Comprensivo de Calificación de los Primeros Recuerdos (*Comprehensive Early Memories Scoring System*, CEMSS-R, Last y Bruhn, 1991), su aproximación a la calificación e interpretación de los recuerdos autobiográficos es muy flexible.[15] Considera a los primeros recuerdos como un fenómeno psicológico complejo cuya explicación puede requerir diferentes modelos teóricos y, en consecuencia, distintos sistemas de calificación. De hecho, propuso el desarrollo de sistemas de calificación elaborados a la medida, o una "boutique" de sistemas, que se basen en aspectos empíricamente observados de los primeros recuerdos de grupos de criterio con el propósito de hacer predicciones específicas. Bruhn y sus colaboradores han obtenido datos prometedores con los sistemas de calificación diseñados para predecir la predisposición a la delincuencia y a la violencia (Davidow y Bruhn, 1990; Tobey y Bruhn, 1992).

El PRT es una técnica que todavía se encuentra en proceso de desarrollo, pues aún no se lleva a cabo el trabajo de normalización, y aunque se han obtenido niveles adecuados de acuerdo entre jueces en varias de las categorías de calificación propuestas

[14] Las instrucciones del PRT también solicitan recuerdos de "una experiencia sexual inapropiada" y de "haber sufrido abuso físico o emocional"; sin embargo, los examinados pueden marcar recuadros para indicar que no han vivido tales experiencias. Esta opción, así como el formato de lápiz y papel del PRT, reducen la posibilidad de obtener el tipo de "recuerdos" sugeridos de abuso que tanto han preocupado a los profesionales de la salud mental como al público en general (véase, por ejemplo, Loftus, 1993).

[15] Una de las técnicas interpretativas que Bruhn más recomienda es el uso de un sumario (o resumen conciso) del recuerdo temprano, un procedimiento que se usa a menudo con el TAT y con otros procedimientos relacionados.

por Bruhn y sus colaboradores, no se cuenta con muchas evidencias empíricas sobre otros tipos de confiabilidad a favor de los primeros recuerdos. Es indudable que la obtención de éstos y otros datos psicométricos sobre el PRT puede ser problemática, como en el caso de otros materiales proyectivos. El sólo acto de categorizar y cuantificar los recuerdos autobiográficos ocasiona una pérdida de información que puede ser de gran importancia para comprender a la persona. No obstante, el procedimiento tiene el potencial para convertirse en una herramienta muy útil para la evaluación de la personalidad, en especial en el contexto de la psicoterapia (véase, por ejemplo, Ritzler, 1993b). Más aún, es probable que muestras sistemáticas de recuerdos autobiográficos sean clínicamente más significativas que otros materiales verbales, como los informes de sueños, las muestras del discurso libre o historias, que han sido útiles para propósitos similares y de maneras semejantes.[16]

TÉCNICAS DE EJECUCIÓN

Una categoría grande y amorfa de técnicas las proyectivas abarca muchas formas de autoexpresión relativamente libre. Una característica común es que todas ellas han sido empleadas como procedimientos terapéuticos y de diagnóstico. Se cree que gracias a las oportunidades de autoexpresión que dichas actividades permiten, el individuo no sólo revela sus problemas emocionales, sino que también puede aliviarlos. En esta categoría se incluyen diversas técnicas de dibujo y de juego, incluyendo el uso dramático de los juguetes. No resulta sorprendente que casi todos esos métodos fueran diseñados para la evaluación de niños, aunque en muchos casos también pueden emplearse con adultos.

Técnicas de dibujo En la búsqueda de señales diagnósticas significativas para la evaluación de la personalidad se ha investigado casi cada medio, técnica o tipo de material subjetivo, aunque se ha dedicado una atención especial al dibujo de la figura humana.[17] Un ejemplo conocido es la Prueba de Dibujo de una Persona de Machover (*Machover Draw-a-Person Test*, D-A-P, 1949), en el que se proporcionan al individuo lápiz y papel y se le pide que "dibuje a una persona". Luego de que termina el primer dibujo, se le pide que dibuje a una persona del sexo opuesto. Mientras el sujeto hace los dibujos, el examinador anota sus comentarios, la secuencia en que dibuja las distintas partes y otros detalles de procedimiento. El dibujo es seguido por una serie de preguntas para obtener información específica sobre edad, escolaridad, ocupación y otros hechos asociados con los personajes dibujados. La indagación puede incluir la petición de que construya una historia acerca de cada personaje.

La interpretación del D-A-P propuesta por Machover es en esencia cualitativa y llena de generalizaciones descomunales que se basan en indicadores simples, como por ejemplo "Las cabezas desproporcionadamente grandes a menudo son dibujadas

[16] El lector puede encontrar en Bruhn (1995b) algunos ejemplos anecdóticos interesantes sobre el valor clínico de los primeros recuerdos, incluida una comparación de los primeros recuerdos tomados de las autobiografías de Nixon y Freud.

[17] En el capítulo 9, analizamos el uso de los dibujos de la figura humana como medida no verbal del funcionamiento cognoscitivo.

por individuos que sufren de daño cerebral orgánico". Aunque se hace referencia a "miles de dibujos" examinados en el contexto clínico y se citan unos cuantos casos seleccionados con propósitos ilustrativos, en la publicación original de la prueba no hay una presentación sistemática de datos. Además, los estudios subsecuentes de validación hechos por otros investigadores no han logrado brindar apoyo a las interpretaciones diagnósticas de Machover (véase, por ejemplo, Klopfer y Taulbee, 1976, pp. 558–561).

Koppitz (1968, 1984) estableció otro método con fundamentos empíricos más firmes para analizar los dibujos de la figura humana hechos por niños y adolescentes. Koppitz cree firmemente en la utilidad clínica del dibujo de la figura humana para evaluar a los niños, por lo que desarrolló y estandarizó dos sistemas objetivos de calificación, utilizando los dibujos realizados por 1 856 niños de escuelas públicas entre los cinco y los 12 años. Uno de los sistemas, basado principalmente en la Prueba de Dibujo de Goodenough-Harris (véase el capítulo 9) y en su propia experiencia clínica, utiliza el dibujo de la figura humana como prueba del desarrollo de la madurez mental. El segundo, derivado en particular del trabajo de Machover, es una prueba proyectiva de las actitudes e intereses interpersonales de los niños y consta de 30 "indicadores emocionales" que distinguen entre los dibujos de niños con y sin problemas emocionales. Estos indicadores rara vez se observaron en los niños normales de la muestra y, a diferencia de las predisposiciones del desarrollo, se supone que no están relacionados con la edad ni con el nivel de maduración; incluyen: (a) *signos de calidad*, como transparencias y sombreado del rostro, (b) *rasgos especiales*, como cabezas pequeñas o figuras grotescas y (c) *omisiones* de algunos elementos esperados, como el cuello o los ojos.

Algunos de los rasgos agregados de los dibujos de la figura humana, como las extravagancias o el número total de "indicadores emocionales", parecen diferenciar a los niños que muestran problemas y de los que están bien ajustados (D. T. Marsh, Linberg y Smeltzer, 1991; Naglieri y Pfeiffer, 1992; Yama, 1990); sin embargo, tanto Koppitz como otros investigadores han advertido en contra del uso de indicadores sencillos como "signos" con propósitos de diagnóstico. Parece existir un consenso respecto a la idea de que los dibujos de figuras humanas sólo pueden proporcionar una idea muy general del nivel de ajuste emocional del niño. Más aún, hasta donde concierne a las aplicaciones de diagnóstico, casi todos los expertos concuerdan en que los dibujos sólo deberían emplearse para generar hipótesis y que al interpretarse deben considerarse otras informaciones sobre el individuo (M. V. Cox, 1993; Knoff, 1993; Tharinger y Stark, 1990).

A pesar de tales advertencias y restricciones, no sólo se mantiene la popularidad de los dibujos de la figura humana, sino que, de hecho, se han incluido tareas adicionales de dibujo. Una de las más populares es la técnica H-T-P (*House-Tree-Person*) que, como lo indica su nombre, requiere que el examinado haga dibujos de una casa, un árbol y una persona (Buck, 1948, 1992). Las características y los rasgos de los dibujos, junto con una extensa indagación que les sigue, suelen utilizarse como fuente de hipótesis sobre las áreas generales de conflicto y preocupación. Una técnica más reciente, que parece tener un potencial desacostumbrado como instrumento clínico, es el Dibujo Dinámico de la Familia (*Kinetic Family Drawing*, KFD, R. C. Burns, 1982; R. C. Burns y Kaufman, 1970, 1972). La prueba pide a los niños que dibujen a todos los miembros

de su familia, incluyéndose, "haciendo algo". El KFD ha generado una gran cantidad de investigaciones. En una revisión reciente de la bibliografía, Handler y Habenicht (1994) afirmaron que a pesar de los problemas metodológicos enfrentados por la investigación hecha sobre esta técnica, hay una serie de descubrimientos alentadores que garantiza su seguimiento con análisis más complejos, como la regresión múltiple. La distancia y el grado de interacción de las figuras del KFD, por ejemplo, parecen estar entre los rasgos psicológicamente más significativos de los dibujos.

Un ejemplo de las tareas incluso más imaginativas que siguen desarrollándose es la técnica de dibujo en colaboración, la cual requiere que la familia o la pareja realicen un dibujo en conjunto en presencia de uno o más terapeutas que observan de cerca la conducta de todos los participantes (G. Smith, 1991). Esta técnica interactiva se inspiró en el KFD y se utiliza sobre todo en el contexto de la terapia familiar.[18]

Técnicas de juego y pruebas con juguetes.[19] En la evaluación proyectiva se han utilizado diversas clases de técnicas de juego y pruebas con juguetes que incluyen objetos como títeres, muñecos y miniaturas. A partir de su origen en la terapia de juego con niños, estos materiales se adaptaron luego para la evaluación diagnóstica de niños y adultos. Los objetos se seleccionan por su valor asociativo, y entre los artículos más usados se encuentran muñecos que representan a adultos y niños de ambos sexos, animales, muebles, figuras de cocina y baño y otros artículos domésticos. El Escenotest, ilustrado en la figura 15.4, que consta de un equipo estandarizado de esta clase y un manual que le acompaña, se publicó en Suiza en la década de los sesenta y acaba de salir a la circulación en los Estados Unidos (Staabs, 1991). Se espera que el juego con estos artículos revele las actitudes del niño hacia su propia familia, así como las rivalidades fraternas, los temores, las agresiones, los conflictos, etc. El examinador toma nota de los artículos que el niño elige y de lo que hace con ellos, así como de sus verbalizaciones, expresiones emocionales y otras formas de conducta abierta.

Cuando se aplica a los niños, el examinador se concreta a presentarles la colección de juguetes para el juego libre; a los adultos les presenta los materiales con instrucciones generales para que realicen alguna tarea de naturaleza no estructurada. Por supuesto, las instrucciones pueden emplearse también con los niños. Con frecuencia, la tarea tiene características de dramatización, como en la disposición de las figuras en un escenario en miniatura; por ejemplo, los materiales empleados para el Escenotest vienen en una caja plana portátil que tiene una tapa que puede utilizarse como "escenario" con las diversas figuras y accesorios (véase la figura 15.4).

En un volumen editado por Schaefer, Gitlin y Sandgrund (1991) se presenta una clasificación de las técnicas de juego para el diagnóstico y la evaluación de los niños. El trabajo no sólo describe las herramientas proyectivas, como el uso de títeres, sino una variedad de escalas de juego para la evaluación de problemas específicos, que van

[18] El lector puede encontrar en Cummings (1986), Hammer (1986) Handler (1996) y Knoff (1990) un análisis subsecuente del uso de los dibujos proyectivos, así como revisiones de varias de las técnicas mencionadas en esta sección.

[19] Véase Krall (1986) para una revisión general de los procedimientos e interpretación del juego proyectivo desde el punto de vista que subraya la teoría psicoanalítica y las perspectivas del desarrollo.

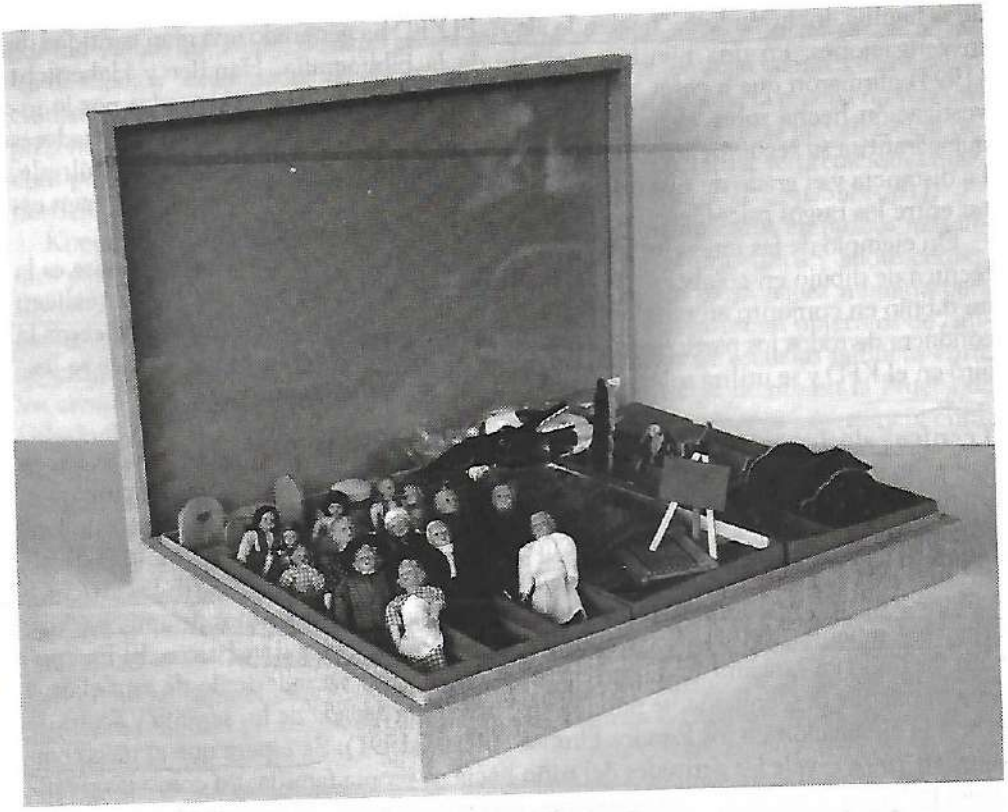


Figura 15.4. Materiales estandarizados para el Escenotest.

(Cortesía de Hogrefe and Huber Publishers. Copyright © 1994 de Hogrefe & Huber Publishers. Reproducido con autorización.)

del autismo a la hiperactividad, y para la evaluación del desarrollo en áreas como la competencia, la destreza y el temperamento infantil. También comprende escalas que pueden usarse en la terapia de juego y para evaluar la interacción entre padres e hijos. Como reconocieran Schaefer y sus colaboradores, muchas de las técnicas que presentan aún se encuentran en las etapas iniciales de desarrollo; no obstante, la diversa colección de aproximaciones que han reunido incluyen algunas medidas ingeniosas y ofrecen métodos de observación formal muy estructurados que pueden ser adecuados para la investigación y muchos instrumentos con orientación clínica.

EVALUACIÓN DE LAS TÉCNICAS PROYECTIVAS

Es claro que las técnicas proyectivas difieren mucho entre sí, algunas parecen más prometedoras que otras gracias a que los descubrimientos empíricos les son más favorables o a que su orientación teórica es más adecuada. En lo que respecta a ciertas técnicas, como el Rorschach, se ha recopilado una cantidad voluminosa de datos,

aunque su interpretación sigue siendo incierta; de otras se sabe muy poco, ya sea por su reciente origen, porque la verificación objetiva es obstaculizada por la naturaleza intrínseca de los instrumentos o por las actitudes de sus exponentes.

Una observación relacionada es que las diferencias entre las técnicas proyectivas y las pruebas estandarizadas no es tan grande ni tan fundamental como puede parecer a primera vista. No sólo en sus propiedades psicométricas, sino también en la naturaleza de la tarea que presentan a los examinados y las formas en que interpretan los resultados, se ha argumentado convincentemente que la diferencia entre las técnicas proyectivas y los inventarios autodescriptivos es más de grado que de calidad (Levy, 1963). Los instrumentos individuales caen en un continuo, y aunque en los extremos sean identificables con facilidad, en el centro resulta evidente el traslape de diversos rasgos.

La evaluación de cada instrumento proyectivo y el intento por resumir la extensa bibliografía correspondiente requeriría un volumen por separado. En este capítulo hemos presentado comentarios críticos sólo en el caso de los instrumentos que presentan características únicas, de naturaleza favorable o desfavorable; sin embargo, pueden plantearse varias preguntas generales acerca de casi todas las técnicas proyectivas, interrogantes que pueden considerarse en una forma resumida.

Rapport y aplicabilidad. Casi todas las técnicas proyectivas representan un medio eficaz para "romper el hielo" durante los primeros contactos entre el clínico y el cliente. La tarea suele ser de por sí interesante y a menudo es entretenida; por lo general, permite que el individuo desvíe la atención de sí mismo y por ende reduce la vergüenza y la tendencia a ponerse a la defensiva. Además, como no hay una respuesta que se considere correcta, no supone mayor amenaza al prestigio del examinado.

Algunas técnicas proyectivas pueden ser útiles con los niños pequeños, los iletrados y las personas con dificultades de lenguaje o defectos del habla. Los medios no verbales serían fácilmente aplicables en esos grupos, y en los dos primeros pueden obtenerse respuestas orales a estímulos visuales y no verbales. Con todos los grupos que presentan limitaciones verbales, las técnicas proyectivas pueden ayudar al examinado a comunicarse con el evaluador y también a que se aclaren algunos aspectos de su comportamiento de los que no había hablado antes.

Falseamiento. En general, los instrumentos proyectivos son menos susceptibles al falseamiento que los inventarios autodescriptivos. El propósito de las técnicas proyectivas es encubierto,[20] e incluso si un individuo tiene cierta complejidad psicológica y está familiarizado con la naturaleza general de algún instrumento, como el Rorschach o el TAT, aun así es poco probable que pueda predecir las intrincadas formas en que las respuestas serán calificadas e interpretadas. El resultado de una serie reciente de estudios en los que se compara la dependencia de una medida proyectiva y una objetiva indica una posible relación inversa entre la validez de facie de una prueba y su susceptibilidad al falseamiento (Bornstein, Rossner, Hill y Stepanian, 1994). Más aún, el

[20] En la medida en que el examinado no está consciente de la naturaleza y el propósito de la evaluación para la que se utilizan esas técnicas, su naturaleza encubierta supone un problema ético (véase el capítulo 18).

examinado pronto queda absorto en la tarea y es poco probable que recurra al falseamiento y las restricciones habituales de la comunicación interpersonal.

Por otro lado, no puede suponerse que las pruebas proyectivas son completamente inmunes al falseamiento. Varios experimentos realizados con el Rorschach, el TAT y otros instrumentos proyectivos han demostrado la existencia de diferencias significativas cuando se dan instrucciones a los examinados para que alteren sus respuestas para crear impresiones favorables o desfavorables o bien cuando se les dan explicaciones que sugieren que ciertas clases de respuestas son más deseables (Masling, 1960). Se cuenta con considerables evidencias experimentales de que las respuestas a las pruebas proyectivas pueden ser alteradas con éxito en ambas direcciones, "para quedar bien" o "para quedar mal", aunque la última puede ser más fácil de lograr. Dichos resultados se han obtenido con varios instrumentos proyectivos, como el Rorschach, el TAT, el Estudio P-F de Rosenzweig y las pruebas de completamiento de frases (Albert, Fox y Kahn, 1980; Exner, 1991; Kaplan y Eron, 1965; Meltzoff, 1951; Netter y Viglione, 1994; Perry y Kinder, 1990; Schwartz, Cohen y Pavlik, 1964). El evaluador experimentado está alerta a las señales de falseamiento, tanto en la naturaleza de las respuestas individuales y los patrones de respuesta como en las inconsistencias con otras fuentes de información sobre el examinado.

El examinador y las variables situacionales. Es evidente que la estandarización de la mayor parte de las técnicas proyectivas resulta inadecuada tanto en la aplicación como en la calificación, o bien que éstas no se usan de forma estandarizada en la práctica clínica. Hay evidencias de que incluso las diferencias sutiles en la formulación de las instrucciones verbales y en las relaciones entre examinado y examinador pueden alterar en buena medida la ejecución en los instrumentos (Baughman, 1951; Exner, 1993; Hamilton y Robertson, 1966; Herron, 1964; Klinger, 1966; Klopfer y Taulbee, 1976). Aunque empleen las mismas instrucciones, algunos examinadores pueden parecer más alentadores e inspirar más tranquilidad mientras que otros resultan amenazadores debido a sus modales y a su apariencia, lo que puede influir en la productividad de respuestas, la tendencia a ponerse a la defensiva, la estereotipia, la imaginación y otras características básicas de la ejecución. A la luz de estas conclusiones, los problemas de aplicación y las condiciones de examinación adquieren una importancia aún mayor que en otros instrumentos psicológicos.

La falta de objetividad es igualmente seria en la calificación y la interpretación. Aunque se han seguido y establecido sistemas objetivos de calificación, el paso final en la evaluación e integración de los datos crudos por lo general depende de la habilidad y la experiencia clínica del examinador. La implicación más inquietante de esta situación es que la interpretación de las puntuaciones a menudo es tan proyectiva para el examinador como lo son los estímulos de la prueba para el examinado. En otras palabras, la interpretación final de las respuestas a las pruebas proyectivas puede revelar más acerca de la orientación teórica, las hipótesis favoritas y la idiosincrasia del examinador que de la dinámica de personalidad del examinado.

Normas. Otra notable deficiencia común a la mayor parte de los instrumentos proyectivos tiene que ver con los datos normativos, que pueden estar ausentes por completo, resultar sumamente inadecuados o basarse en poblaciones descritas con ambigüe-

dad. Al carecer de normas objetivas adecuadas, el clínico se apoya en su "experiencia clínica general" para interpretar la ejecución en las pruebas proyectivas, pero ese marco de referencia está sometido a todas las distorsiones de la memoria que en sí mismas son reflejos de los sesgos teóricos, las preconcepciones y otras idiosincrasias del clínico; más aún, sus contactos pueden haberse limitado a personas atípicas en su nivel educativo, socioeconómico, en la proporción de sexos, la distribución de edad y otras características importantes. Al menos en un aspecto es casi seguro que la experiencia del clínico produce una imagen errónea, ya que al tratar principalmente con individuos que sufren perturbaciones o patologías puede no estar lo bastante familiarizado con las reacciones de la gente normal en las pruebas. Las normas del Rorschach obtenidas por Exner representan un esfuerzo por corregir algunas de las carencias más evidentes a ese respecto.

A menudo, la interpretación de la ejecución en las pruebas proyectivas incluye normas de subgrupo, que pueden ser de naturaleza objetiva o subjetiva, y que pueden llevar a una interpretación errónea a menos que los subgrupos se hubieran igualado en otros aspectos; por ejemplo, si los individuos esquizofrénicos y normales de los que se derivaron las normas también diferían en su nivel de escolaridad, las diferencias observadas entre la ejecución de ambos puede ser resultado de las diferencias educativas más que de la esquizofrenia. Los errores sistemáticos o constantes también pueden aparecer en la comparación de varios síndromes psiquiátricos; por ejemplo, hay evidencias de que en algunos grupos étnicos y en los individuos más jóvenes los clínicos tienden a diagnosticar esquizofrenia en lugar de trastorno bipolar; de igual manera, se ha informado que el trastorno de conversión es más común en individuos de nivel socioeconómico bajo (American Psychiatric Association, 1994).

Confiabilidad.[21] La *confiabilidad del calificador* adquiere particular relevancia al considerar la naturaleza especial de los procedimientos de puntuación y de lo inadecuado de los datos normativos en las evaluaciones proyectivas.[22] En el caso de las técnicas proyectivas, una medida adecuada de la confiabilidad del calificador no sólo debería incluir las puntuaciones preliminares más objetivas, sino también las etapas finales, cuando se hace la integración y la interpretación; por ejemplo, no basta demostrar que en las calificaciones hechas por los examinadores que dominan el mismo sistema del Rorschach hay una estrecha semejanza en la anotación de características como un todo, los detalles desacostumbrados o las respuestas de color. A diferencia de lo que sucede en otras pruebas psicológicas, en un instrumento como el Rorschach esas medidas cuantitativas crudas no pueden interpretarse directamente de una tabla de normas. La confiabilidad interpretativa del calificador tiene que ver con el grado en que diferentes evaluadores atribuyen las mismas características de personalidad al examinado en función de sus interpretaciones de un registro idéntico. Se han realizado muy pocos estudios adecuados de esta clase de confiabilidad del calificador con las técnicas proyectivas. Algunos investigadores han hecho evidente la existencia de notables diver-

[21] El lector puede encontrar en C. P. Smith (1992) una revisión detallada de los temas de confiabilidad que tienen que ver con los tests de apercepción temática, en la que se hace referencia especial a las variables motivacionales.

[22] En reconocimiento a la importancia de asegurar un buen nivel de objetividad en las prácticas de calificación, desde 1991 el *Journal of Personality Assessment* exige a los autores de los artículos que informan acerca de la investigación sobre el Rorschach que proporcionen evidencias de al menos 80 por ciento de acuerdo entre dos o más calificadores para todas las categorías importantes de calificación.

gencias en las interpretaciones hechas por usuarios razonablemente bien calificados; una parte fundamental de la ambigüedad de los resultados obedece a la contribución desconocida de la capacidad del que hace la interpretación. Sea elevada o baja, la confiabilidad del calificador no puede generalizarse directamente a otros calificadores que difieren de los que participaron en la investigación particular. De hecho, una de las principales causas de la popularidad de los sistemas computarizados de interpretación de pruebas como el Rorschach es la uniformidad de sus resultados en lo que atañe a la interpretación.

En el caso de los instrumentos proyectivos, los intentos por medir otras clases de confiabilidad de la prueba han corrido con peor suerte. Los coeficientes de *consistencia interna*, si acaso se calculan, por lo general son bajos. Se argumenta que en instrumentos como el Rorschach, el TAT y el Estudio P-F de Rosenzweig las distintas tarjetas o reactivos no son comparables y que, por consecuencia, no pueden usarse para calcular la confiabilidad de división por mitades. De hecho, cada reactivo de estos instrumentos se diseñó para medir variables distintas, y la tendencia de las respuestas en los reactivos sucesivos a menudo se considera importante en la interpretación. J. W. Atkinson (1981; J. W. Atkinson y Birch, 1978, pp. 370–374) demostró mediante simulación por computadora que, con un instrumento similar al TAT, es posible obtener una elevada validez de constructo de las puntuaciones totales (por ejemplo, .90) cuando la consistencia interna es muy baja (digamos, .07). Además, observó que el individuo no responde a cada tarjeta sucesiva de manera independiente, sino que lo hace en una corriente continua de actividad que refleja el aumento y la disminución en la fuerza relativa de distintas tendencias de conducta, y que la expresión de una tendencia en la actividad reduce su fuerza. Así, el tiempo que el individuo dedica a describir, por ejemplo, las actividades motivadas por el logro en respuesta a diferentes tarjetas es una función del efecto acumulativo de responder a tarjetas sucesivas y los incentivos diferenciales por el logro y otros motivos que compiten en las tarjetas individuales. En vista de los diversos argumentos en contra de la posibilidad de aplicar las medidas de confiabilidad por consistencia interna a las técnicas proyectivas, una solución consiste en elaborar *formas paralelas* que sean en efecto comparables a lo hecho en la técnica de manchas de tinta de Holtzman.

La *confiabilidad del retest* también plantea problemas especiales. Con intervalos largos pueden ocurrir genuinos cambios en la personalidad que la prueba debería detectar; pero, con intervalos cortos, el retest puede mostrar apenas algo más que el recuerdo de las respuestas originales. Cuando en el retest con el TAT los investigadores pedían a los examinados que escribieran historias diferentes para determinar si recurrían los mismos temas, la mayor parte de las variables calificadas producía correlaciones insignificantes en el retest (Lindzey y Herman, 1955). También es importante advertir que muchas puntuaciones derivadas de las técnicas proyectivas se basan en muestras de respuestas muy inadecuadas; por ejemplo, en un protocolo individual del Rorschach el número de respuestas correspondientes a categorías como movimiento animal, movimiento humano, sombreado, color, detalles desacostumbrados y cosas por el estilo puede ser insuficiente y producir índices muy poco confiables, de ahí que se esperen grandes variaciones al azar. Las proporciones y los porcentajes calculados con esas medidas poco confiables son todavía más inestables que las propias medidas individuales (Cronbach, 1949, pp. 411–412).

Validez. En cualquier prueba, el aspecto fundamental es la validez. Muchos estudios de validación de los instrumentos proyectivos se han centrado en la validación concurrente relacionada con un criterio. La mayor parte ha comparado la ejecución de grupos contrastados, como grupos ocupacionales o de diagnóstico; sin embargo, como dijimos en relación con las normas, a menudo estos grupos difieren también en otros aspectos, como la edad o la escolaridad. Otras investigaciones de validez concurrente han empleado en esencia una técnica de emparejamiento, en la que las descripciones de la personalidad derivadas de los registros de la prueba se comparan con las descripciones o datos de la misma persona tomados de las historias de casos, entrevistas psiquiátricas o registros conductuales a largo plazo. Muy pocos estudios han investigado la validez predictiva sobre criterios como el éxito en la capacitación especializada, el desempeño en el trabajo o la respuesta a la psicoterapia. Por otro lado, se ha observado una tendencia creciente a investigar la validez de constructo de las técnicas proyectivas al someter a prueba hipótesis específicas que sustentan el uso y la interpretación de cada instrumento.

La mayor parte de los estudios de validación publicados sobre las técnicas proyectivas no son concluyentes por las deficiencias de procedimiento en los controles experimentales, el análisis estadístico o ambos. Algunas deficiencias metodológicas pueden tener el efecto de producir *evidencia espuria de validez* donde no existe. Un ejemplo de ello es la contaminación de los criterios o los datos de la prueba; así, los jueces del criterio pueden haber tenido cierto conocimiento acerca de la ejecución del examinado en la prueba. De modo similar, el examinador puede haber obtenido señales sobre las características del examinado al conversar con él durante la aplicación, del material de la historia del caso o merced a otras fuentes. El control acostumbrado para el último tipo de contaminación en los estudios de validación consiste en utilizar análisis ciegos, en el que el registro de la prueba es interpretado por un calificador que no haya tenido contacto con el examinado y que no tiene más información que la contenida en el protocolo del instrumento; sin embargo, los clínicos argumentan que no es una forma natural de interpretar las respuestas a las pruebas proyectivas y que no corresponden a la manera en que esos instrumentos suelen emplearse en la práctica clínica.

Otra fuente común de datos de validez espuria es la falta de validación cruzada (Kinslinger, 1966). Debido al gran número de signos potenciales de diagnóstico o elementos calificables que pueden derivarse de la mayor parte de las técnicas proyectivas, resulta muy sencillo encontrar un conjunto de signos que por azar permita hacer una diferenciación significativa entre grupos de criterio; no obstante, la validez de esta clave de puntuación caerá a cero cuando se aplique a nuevas muestras.

Una forma más sutil de error se encuentra en el estereotipo de precisión. Ciertas afirmaciones descriptivas, como las que es posible encontrar en el protocolo del Rorschach, pueden aplicarse a toda la gente, a los hombres jóvenes, a los pacientes hospitalizados o a cualquier categoría que se haya muestreado para la investigación particular.[23] En consecuencia, el acuerdo entre el criterio y los datos de la prueba con respecto a

[23] El uso de estas afirmaciones generalmente aplicables constituye un ejemplo del "efecto Barnum" citado en el capítulo 17 (Dunnette, 1957; Meehl, 1956). Véase en Klopfer (1983, pp. 510–514) para un estudio bien equilibrado de la gran cantidad de investigación realizada sobre este efecto.

dichas afirmaciones producirá una falsa impresión de validez. Se requiere una supervisión de este error, como una medición del acuerdo entre la evaluación que hace la prueba de un individuo y la evaluación criterio de otro individuo en la misma categoría. Esta medida indicaría el monto del acuerdo espurio que resulta del estereotipo de precisión en las condiciones de la investigación particular (véase, por ejemplo, L. H. Silverman, 1959).

Otra fuente común de error, que surge de la confianza en la experiencia clínica para la validación de los signos diagnósticos, es lo que Chapman (1967) llamó "validación ilusoria", un fenómeno que puede explicar que continúe el uso de instrumentos y sistemas de signos de diagnóstico para los cuales los datos de validez empírica son más que nada negativos. En una serie clásica de experimentos diseñados para probar este fenómeno, Chapman y Chapman (1967) presentaron a estudiantes universitarios un conjunto de dibujos de figuras humanas similares a los obtenidos con la Prueba de dibujo de una persona de Machover (D-A-P). Los resultados mostraron que los participantes respondían en términos de los estereotipos populares preestablecidos aun cuando esas asociaciones no fueran apoyadas por los datos que se les habían presentado durante su "experiencia de aprendizaje"; por ejemplo, habían escuchado que los ojos atípicos se asociaban con la suspicacia, la cabeza grande con la preocupación por la inteligencia y los hombros amplios con la preocupación por la masculinidad. Las interpretaciones no sólo no se relacionaron con las asociaciones empíricas que los participantes habían "estudiado", sino que en otros experimentos las asociaciones culturales estereotipadas demostraron ser resistentes al cambio en condiciones intensivas de capacitación pensadas para establecer asociaciones contrarias. En otras palabras, el individuo retenía sus expectativas incluso después de ser expuesto a observaciones contradictorias.

La validación ilusoria es un ejemplo especial del mecanismo que está detrás de la supervivencia de la superstición. Tendemos a percibir y recordar lo que concuerde con nuestras expectativas y a ignorar y olvidar lo que sea contrario a ellas. Este mecanismo puede interferir con el descubrimiento y uso de signos diagnósticos válidos en el curso de la observación clínica hecha por clínicos que están muy comprometidos con un sistema particular de diagnóstico. La investigación original de Chapman y Chapman con el D-A-P ha sido corroborada en estudios similares con el Rorschach y con el Formulario de Frases Incompletas (Chapman y Chapman, 1969; Golding y Rorer, 1972; Starr y Katkin, 1969).

Por otro lado, debe advertirse que ciertas insuficiencias del diseño experimental pueden *subestimar la validez* del instrumento de diagnóstico; por ejemplo, es muy reconocido que las categorías psiquiátricas tradicionales, como la esquizofrenia o el trastorno de despersonalización, representan clasificaciones gruesas de los trastornos realmente manifestados por los pacientes. De ahí que si esas categorías de diagnóstico se utilizan como el único criterio para supervisar la validez de una prueba de personalidad, los resultados negativos no son concluyentes. De igual modo, el fracaso para predecir los criterios ocupacionales puede reflejar apenas la ignorancia del examinador de los rasgos requeridos para el empleo considerado. Cuando se utilizan estos criterios, es posible que el instrumento proyectivo sea una medida válida de los rasgos de personalidad que pretende evaluar, pero que esos rasgos sean irrelevantes para el éxito en las situaciones de criterio elegidas.

Cada vez son más los usuarios de las pruebas que insisten en la importancia de principios holistas e integradores en la evaluación de la personalidad, como el uso de patrones de configuración y variables contextuales. Muchos de ellos han criticado los intentos continuos por validar indicadores únicos, calificaciones aisladas o "signos" diagnósticos derivados de las técnicas proyectivas. Una plétora de descubrimientos contradictorios acerca de muchas de las técnicas proyectivas que los clínicos consideran más útiles ilustra el hecho de que las correlaciones insignificantes puedan ser resultado de la incapacidad para considerar patrones complejos de relaciones entre variables de personalidad; por ejemplo, la relación hipotetizada entre la agresión fantaseada (revelada en el TAT) y la agresión en la conducta abierta no es simple, ya que dependiendo de otras características de personalidad, como el nivel de ansiedad o el temor al castigo, una elevada agresión fantaseada puede asociarse con una agresión abierta elevada o baja (R. Harrison, 1965; Mussen y Naylor, 1954).

La falta de correlación significativa entre la expresión de la agresión en las historias del TAT y la que se observa en la conducta abierta en una muestra aleatoria de individuos es congruente con las expectativas, ya que la relación puede ser positiva en algunas personas y negativa en otras; sin embargo, es obvio que la falta de correlación también es congruente con la hipótesis de que la prueba no tiene validez alguna para detectar las tendencias agresivas. Es claro que se requieren más estudios que utilicen diseños experimentales complejos que permitan un análisis de las condiciones en las cuales puede aplicarse cada suposición.

La hipótesis proyectiva. La suposición tradicional respecto a las técnicas proyectivas es que las respuestas del individuo a los estímulos que se le presentan reflejan atributos importantes y relativamente duraderos de su personalidad. Aunque es claro que las respuestas a las técnicas proyectivas sólo pueden reflejar los estilos de respuesta y rasgos constantes de los individuos, una cantidad importante de investigaciones indica que muchos otros factores pueden influir en estas respuestas. Cuando se ha evaluado la confiabilidad del retest, es frecuente observar notables cambios temporales que indican la operación de un error aleatorio importante. Diversos estudios experimentales que demuestran el efecto que tienen factores como el hambre, la privación de sueño, las drogas, la ansiedad y la frustración sobre las respuestas a los instrumentos proyectivos ofrecen evidencias más directas sobre la susceptibilidad de estas respuestas a los estados temporales. De igual modo, en las respuestas se han encontrado diferencias importantes relacionadas con el conjunto de instrucciones, las características del examinador y la percepción que el individuo tiene de la situación de evaluación. Los factores de habilidad, y en particular la verbal, afectan sin duda las puntuaciones de casi todas las pruebas proyectivas. Al considerar esos descubrimientos, resulta claro que las pruebas proyectivas sólo pueden interpretarse cuando el examinador tiene una vasta información sobre las circunstancias en las que se obtuvieron y sobre las aptitudes y las experiencias del examinado.

Desde otro punto de vista, se ha cuestionado la ventaja de utilizar estímulos ambiguos o no estructurados (Epstein, 1966), ya que no sólo son ambiguos para el examinado, sino también para el examinador, por lo que tienden a incrementar la ambigüedad en la interpretación de las respuestas. Por otro lado, cuando se trabaja con estímulos es-

tructurados es posible seleccionar los que sean relevantes para las características de personalidad que se desea evaluar y variar la naturaleza de los estímulos para explorar plenamente una determinada dimensión de personalidad. Semejante procedimiento permite hacer interpretaciones más claras de la ejecución en una prueba de lo que es posible con la aproximación forzada de los estímulos no estructurados. También hay evidencias en contra de la suposición común de que entre menos estructurados sean los estímulos, más probable es que provoquen la proyección y que toquen las capas más "profundas" de la personalidad (Klopfer y Taulbee, 1976; Murstein, 1963). En realidad, entre la ambigüedad y la proyección parece haber una relación no lineal en la que un grado intermedio de ambigüedad representa una condición óptima para propósitos de proyección.

También se ha cuestionado la suposición de que la fantasía, como la que provocan técnicas proyectivas como el TAT, revela disposiciones motivacionales encubiertas; por ejemplo, en un estudio longitudinal realizado durante 20 años del TAT sobre la fantasía y la conducta abierta relevante, se descubrió que las actividades de los adolescentes permitían hacer una predicción mejor de la imaginería del adulto en el TAT de lo que podía predecir la imaginería del adolescente en la misma prueba respecto de las actividades del adulto (McClelland, 1966; Skolnick, 1966). En un caso, los individuos que habían tenido movilidad social ascendente obtuvieron puntuaciones más altas en las necesidades de logro cuando eran adultos, pero los que habían obtenido puntuaciones mayores en la necesidad de logro durante la adolescencia no estaban entre los que mostraban movilidad social ascendente.

Descubrimientos como éstos invierten la relación implicada en la lógica tradicional de las técnicas proyectivas; pueden explicarse si consideramos que las respuestas al TAT no son expresiones proyectivas directas de los motivos, sino muestras de los pensamientos del individuo, que a su vez pueden haber sido influidas por sus acciones previas. Los individuos que han logrado más y los que han estado más expuestos a modelos orientados al logro en su desarrollo tienden a percibir más temas de logro en las imágenes ambiguas.

En suma, muchas investigaciones han generado dudas acerca de los diversos aspectos de las hipótesis proyectivas y se cuenta con evidencias de que diversas explicaciones alternativas pueden dar cuenta de las respuestas del individuo a los estímulos ambiguos o no estructurados de la prueba.

Las técnicas proyectivas como instrumentos psicométricos. Está claro que muchas técnicas proyectivas son deficientes cuando se evalúan de acuerdo con los estándares psicométricos, lo que se vuelve evidente a partir de los datos resumidos en las secciones precedentes sobre la estandarización de los procedimientos de aplicación y calificación, lo adecuado de las normas, la confiabilidad y la validez. Resulta verdaderamente impresionante la cantidad de estudios publicados que no han logrado demostrar la validez de técnicas como el TAT y el D-A-P, incluso después de descartar los errores metodológicos de muchos de estos estudios. Empero, después de varias décadas de resultados contradictorios, el uso de las técnicas proyectivas permanece sustancialmente sin cambios y de hecho puede estar experimentando un fuerte crecimiento. En palabras de un crítico: "Aún existen clínicos entusiastas y estadísticos dubitativos" (Adcock, 1965, p. 533).

Tal vez podamos entender esta aparente contradicción si reconocemos que las técnicas proyectivas no son verdaderas pruebas, salvo algunas excepciones entre las que se encuentran la técnica de manchas de tinta de Holtzman, el trabajo realizado por Exner con el Rorschach, algunas adaptaciones del TAT, ciertas pruebas de completamiento de oraciones y el Estudio P-F de Rosenzweig. Es indudable que entre las muchas técnicas proyectivas que no revisamos en este capítulo se encontrarán otros ejemplos de cuasipruebas; sin embargo, incluso en el caso de tales instrumentos se necesitan muchos datos más de validez para especificar la naturaleza de los constructos medidos por sus puntuaciones y muchos más datos normativos sobre poblaciones definidas con claridad. Tenemos entonces que, aun cuando se acercan a los estándares de otras técnicas proyectivas, la mayor parte de esos instrumentos no está preparada para el uso operacional rutinario para ayudar a tomar decisiones y hacer predicciones acerca de la gente.

Las técnicas proyectivas como instrumentos clínicos. En lugar de considerarse y evaluarse como instrumentos psicométricos o *tests* en el estricto sentido del término, la mayor parte de los instrumentos proyectivos ha llegado a considerarse como herramientas clínicas, de modo que en las manos de un clínico experimentado pueden servir como ayuda complementaria en la entrevista cualitativa. Su valor como herramienta clínica es proporcional a la habilidad del clínico, por lo que la evaluación no puede hacerse independientemente del usuario. En tal sentido, pueden revelarse como inadecuados los intentos de evaluación en términos de los procedimientos psicométricos comunes. Por la misma razón, el uso de sistemas elaborados de calificación que arrojan puntuaciones cuantitativas no sólo es antieconómico, sino que también puede ser engañoso. Estos sistemas de calificación proporcionan a los resultados un aspecto ilusorio de objetividad y pueden crear la impresión injustificada de que la técnica puede considerarse como una prueba. Es más probable que el valor especial que las técnicas proyectivas pueden tener surja cuando se interpretan mediante procedimientos clínicos cualitativos que cuando se califican cuantitativamente y se interpretan como si fueran instrumentos psicométricos objetivos.

Tomando un concepto propio de la teoría de la información, Cronbach y Gleser (1965) caracterizaron la entrevista y las técnicas proyectivas como procedimientos de "banda ancha". La amplitud de la banda, o la cobertura, se obtiene a costa de una menor fidelidad o confiabilidad de la información. Las pruebas psicométricamente objetivas producen una banda de información más estrecha con un mayor nivel de seguridad. En contraste, las técnicas proyectivas y de entrevista proporcionan un margen mucho más amplio de información con una seguridad menor. Más aún, los datos obtenidos con cualquier técnica proyectiva varían de un individuo a otro; por ejemplo, las respuestas de un individuo al TAT pueden decirnos mucho acerca de su agresión y poco o nada acerca de su creatividad o necesidad de logro; el registro de otra persona puede permitir una evaluación más minuciosa del grado de creatividad y la fuerza de la necesidad de logro, pero revelar muy poco acerca de su agresión. Esta falta de uniformidad en la información proporcionada en los casos individuales ayuda a explicar la baja validez que se encuentra al analizar las respuestas a los instrumentos proyectivos para cualquier rasgo único en un grupo de personas.

Es interesante mencionar que una desigualdad similar caracteriza las interpretaciones que hacen los clínicos de los registros individuales. Así, en un estudio pionero sobre la validez del TAT Henry y Farley (1959, p. 22) concluyeron que:

> No hay una sola forma correcta de emplear la interpretación del TAT. Hubo poco acuerdo entre jueces en los reactivos, pero cada uno tomaba las suficientes decisiones "correctas" para producir una cifra de acuerdo muy significativa. Los jueces podían llegar esencialmente a las mismas implicaciones interpretativas en el informe de la prueba, pero por rutas muy diferentes, o bien podían diferir en su habilidad para utilizar las predicciones del TAT en áreas distintas [...] o con sujetos diferentes.

Los psicólogos prestan cada vez más atención a la naturaleza del juicio clínico con el que pueden usarse los datos proyectivos y de la entrevista para tomar decisiones sobre clientes (véase el capítulo 17). En este proceso, se forman inductivamente los constructos o las categorías en los que se organiza la información mediante el examen de la combinación particular de datos en cada caso. La función especial de los clínicos consiste en hacer predicciones a partir de combinaciones únicas o raras de acontecimientos sobre las cuales no resulta práctico preparar alguna tabla o ecuación estadística. Al crear nuevos constructos ajustados al individuo, el clínico puede hacer predicciones partiendo de combinaciones de acontecimientos que nunca antes ha enfrentado el sujeto. Al hacerlo, puede tener en cuenta la diversa importancia que sucesos similares tienen para diferentes individuos. Estas predicciones clínicas resultan de ayuda toda vez que no se aceptan como últimas, sino que constantemente se contrastan con la información obtenida mediante la indagación subsecuente, las respuestas a las pruebas, la reacción a la terapia u otras conductas del cliente. De la naturaleza de la entrevista y de las técnicas proyectivas se sigue que las decisiones no deberían basarse en ningún dato simple o puntuación obtenida de estas fuentes. Las técnicas funcionan mejor en decisiones seriadas al sugerir guías para posteriores exploraciones o hipótesis sobre el individuo que luego se sometan a verificación.

Otras técnicas de evaluación

Aunque los inventarios autodescriptivos y las técnicas proyectivas que revisamos en los capítulos precedentes son los instrumentos mejor conocidos y de mayor uso para la valoración de la personalidad, en nuestros días se exploran muchos instrumentos de evaluación, por lo que se espera que de esa diversidad de aproximaciones surjan técnicas que estimulen el progreso en nuevas direcciones. En este capítulo veremos en particular ciertos procedimientos empleados como técnicas de investigación, aunque algunos se utilizan también como herramientas complementarias de evaluación en contextos aplicados como la consejería o la psicología organizacional. Las técnicas que explicaremos representan diversos planteamientos, y la clasificación de algunas de ellas es difícil porque evalúan constructos que abarcan los campos de las habilidades y de la personalidad. Tres de las categorías principales incluyen mediciones de estilos cognoscitivos y tipos de personalidad, pruebas situacionales y técnicas diseñadas para evaluar el autoconcepto y los constructos personales. También nos detendremos en el uso de técnicas que no se consideran pruebas (como las observaciones naturalistas, las entrevistas, las valoraciones y el análisis de datos de la historia vital), lo que aumenta la profundidad de nuestra revisión.

MEDIDAS DE ESTILOS Y TIPOS

Aunque en este libro como en cualquier otra fuente se analizan por separado la evaluación de la habilidad y de la personalidad, en el uso actual de las pruebas y en la interpretación de sus puntuaciones es imposible separar los campos cognoscitivo y

afectivo.[1] Las muestras conductuales que componen las pruebas psicológicas son secciones cruzadas del repertorio conductual del individuo y, como tales, contienen a la vez información sobre todos sus aspectos; por ejemplo, en el capítulo 11, revisamos las evidencias sobre el consenso creciente respecto de que las aptitudes no pueden investigarse independientemente de las variables afectivas porque la ejecución en las pruebas de habilidad depende sin duda de la pulsión de logro, la persistencia, los valores y otros factores similares. De igual modo, en el capítulo 13 analizamos la idea de que los individuos difieren en la consistencia que su comportamiento muestra en diversas situaciones como función de la forma en que perciben y categorizan las situaciones, lo que a su vez depende de su experiencia y del aprendizaje anterior. Además, en los capítulos de la tercera y cuarta partes nos ocupamos de algunas medidas que pretenden evaluar la habilidad o la personalidad y que incluyen aspectos relacionados con el funcionamiento afectivo o cognoscitivo respectivamente.

Hay diversas maneras de manejar la multiplicidad de factores inherentes al comportamiento humano. Por ejemplo, podemos examinar la correlación entre las medidas de rasgos como la ansiedad y la habilidad para resolver problemas (véase, entre otros, Zeidner, 1995), o bien podemos emplear técnicas multivariadas, como el análisis factorial y el escalamiento multidimensional, para separar los componentes en un conjunto de datos conductuales (Jones y Sabers, 1992). También es posible desarrollar los esquemas estructurales que tienen cabida para diversas variables y representan sus interrelaciones topográficamente. El modelo jerárquico de la organización de habilidades que preparó Vernon (véase el capítulo 11), la representación hexagonal que hizo Holland de los temas ocupacionales (véase el capítulo 14) y también el modelo interpersonal circunflejo (véase, por ejemplo, Hofstee *et al.*, 1992) son ejemplos de esta clase de arreglo en los campos simples del comportamiento.[2] Otra aproximación consiste en utilizar categorías complejas y que incluyan elementos tanto afectivos como cognoscitivos. La inteligencia emocional (Salovey y Mayer, 1990; Mayer y Salovey, 1993) es uno de esos constructos que en los últimos tiempos ha recibido considerable atención (véase, por ejemplo, Goleman, 1995). Los estilos cognoscitivos y los tipos de personalidad, que veremos en las dos secciones siguientes, también son ejemplo de este acercamiento y representan la tentativa por abarcar las diferencias cualitativas en los patrones o configuraciones del comportamiento humano.

Estilos cognoscitivos. La expresión *estilos cognoscitivos* se refiere a los modelos preferidos y característicos del individuo para percibir, recordar, pensar y resolver problemas (Messick *et al.*, 1976). Se consideran estilos conductuales característicos que abarcan habilidades y personalidad y se manifiestan en diversas actividades y medios. Se ha acumulado una extensa bibliografía de la investigación realizada sobre los estilos cognoscitivos y otros conceptos relacionados como los estilos para aprender y para pensar (Brodzinsky, 1982; Furnham, 1995; Globerson y Zelniker, 1989; Goldstein y

[1] Cada vez se reconoce más la necesidad de integrar esos dominios en la teoría y la investigación psicológica general, lo mismo que en la práctica. Véase, por ejemplo, los dos volúmenes recientes dedicados a la interfase entre personalidad e inteligencia, editados por Saklofske y Zeidner (1995) y por Sternberg y Ruzgis (1994).

[2] La representación esférica de los intereses vocacionales propuesta por Tracey y Rounds (1996), mencionada en el capítulo 14, es el ejemplo de un esquema más complejo que comprende dimensiones múltiples.

Blackman, 1978a, 1978b; Grigorenko y Sternberg, 1995; Jonassen y Grabowski, 1993; Kogan, 1976; Messer, 1976; Sternberg, 1994; Witkin y Goodenough, 1981).

En el área de las funciones perceptuales puede encontrarse una de las principales fuentes de diferenciación de los estilos cognoscitivos. La bibliografía experimental ha demostrado la existencia de relaciones significativas entre las características actitudinales, motivacionales o emocionales del individuo y su desempeño en tareas cognoscitivas o perceptuales. Por otro lado, también hay que reconocer que algunas técnicas proyectivas, de manera notable el Rorschach, pueden considerarse esencialmente como pruebas perceptuales (véase, por ejemplo, Blatt, 1990).

Uno de los factores identificados en los primeros análisis factoriales de la percepción que ha demostrado ser particularmente fructífero en la investigación de la personalidad es la flexibilidad del cierre (Pemberton, 1952; Thurstone, 1944). En general, las pruebas que evalúan este factor requieren la identificación de una figura rodeada de detalles distractores y confusos. En la figura 16.1 se muestran dos reactivos de un instrumento con una carga elevada en este factor (figuras Gottschaldt). En varios de los primeros estudios presentados se informaba sobre datos que indicaban la posible relación entre este factor perceptual y los rasgos de personalidad; por ejemplo, en una investigación se descubrió que quienes destacaban en la flexibilidad de cierre se asignaban mejores calificaciones en descripciones de rasgos como la reserva social, la independencia de las opiniones ajenas, el carácter analítico, el interés en los problemas teóricos y científicos y el disgusto por la sistematización y la rutina (Pemberton, 1952). En los años subsecuentes los investigadores han utilizado adaptaciones de las figuras Gottschaldt en la investigación de la conducta cognoscitiva y no cognoscitiva.

Al acercarse al problema desde un punto de vista distinto, en el estudio a largo plazo de la orientación perceptual en el espacio Witkin y sus colaboradores (Witkin *et al.*,

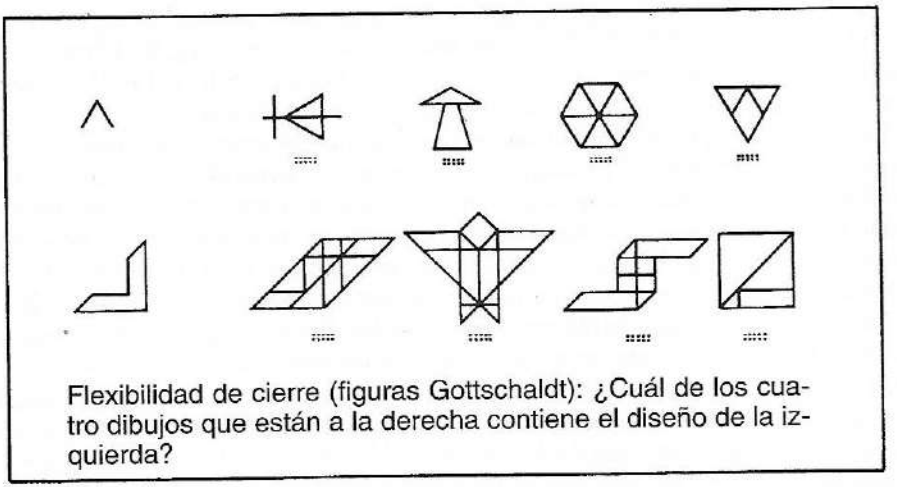


Figura 16.1. Reactivos que ilustran las tareas perceptuales empleadas en la evaluación de la personalidad.

(Tomado de Thurstone, 1950, p. 7.)

1954/1972) identificaron como variable importante la habilidad de resistir a la influencia desorganizada de señales contextuales de conflicto. Los investigadores demostraron que los individuos difieren considerablemente en su "dependencia del campo" o en el grado en que su percepción de la vertical está influida por el campo visual que lo rodea mediante el uso de diversas pruebas que emplean una vara y un marco que pueden moverse independientemente, una silla y una habitación inclinadas. La confiabilidad fue elevada tanto en el retest como en pares y nones; además, casi todas las correlaciones entre las diferentes pruebas de orientación espacial fueron significativas. Tenemos entonces que se cuenta con una cantidad considerable de datos que indican que la dependencia del campo es un rasgo consistente, relativamente estable y con cierto grado de generalidad.

De mayor interés aún son las correlaciones significativas entre las pruebas de orientación y el Test de Figuras Encajadas (*Embedded Figures Test*) (similar a las figuras Gottschaldt ilustradas en la figura 16.1) que se considera como una medida de dependencia del campo en una situación puramente visual de lápiz y papel. A medida que aumentaba la investigación llegó a considerarse a la dependencia o independencia del campo como el componente perceptual de una dimensión más amplia de la personalidad —conocida como estilo cognoscitivo global a diferencia del articulado— o como diferenciación psicológica (Witkin, Dik, Faterson, Goodenough y Karp, 1962/1974). Hay evidencias que indican que este estilo cognoscitivo se mantiene considerablemente estable de la niñez a la juventud y que se relaciona con variables de personalidad como el liderazgo (Weissenberg y Gruenfeld, 1966) y la conformidad social (Witkin *et al.*, 1974) entre otras (véase, por ejemplo, Jonassen y Grabowski, 1993, capítulo 7).

Resulta impresionante el alcance y la diversidad de la investigación sobre la dependencia de campo. Va de las relaciones interpersonales (Witkin y Goodenough, 1977) al aprendizaje y la memoria (D. R. Goodenough, 1976), el aprovechamiento en matemáticas (Vaidya y Chansky, 1980), la elección del campo de estudios en la universidad y el posgrado (Raskin, 1985), las diferencias entre culturas (J. W. Berry, 1976) y las preferencias por ambientes laborales (Wooten, Barner y Silver, 1994). Al revisar estos estudios se hacen evidentes algunas relaciones interesantes, como el hecho de que los que son independientes del campo tienden a aproximarse al aprendizaje de manera activa y "participativa", mientras que quienes son dependientes se acercan más como "espectadores"; por ejemplo, una investigación reciente que comparó los resultados en los exámenes de opción múltiple y en las evaluaciones basadas en el rendimiento indica que la última favorece a los estudiantes independientes del campo (Lu y Suen, 1995). Por otro lado, en las situaciones interpersonales a los dependientes del campo les resulta más fácil llevarse bien con los demás, ya que suelen prestar más atención a las señales sociales, responden más a la conducta de los otros y son emocionalmente más abiertos que los independientes del campo. Tenemos entonces que ninguno de los extremos del continuo de dependencia–independencia del campo es necesaria o uniformemente favorable o desfavorable, sino que el valor de las desviaciones en cualquier dirección depende de las exigencias de las situaciones particulares.

Buena parte de la investigación mencionada ha utilizado el Test de Figuras Encajadas cuya aplicación es relativamente sencilla. Se dispone de versiones para adultos, niños y preescolares, así como para aplicación grupal (Coates, 1972; Witkin, Oltman, Raskin y Karp, 1971). En la figura 16.2 se presentan dos reactivos del instrumento. En las for-

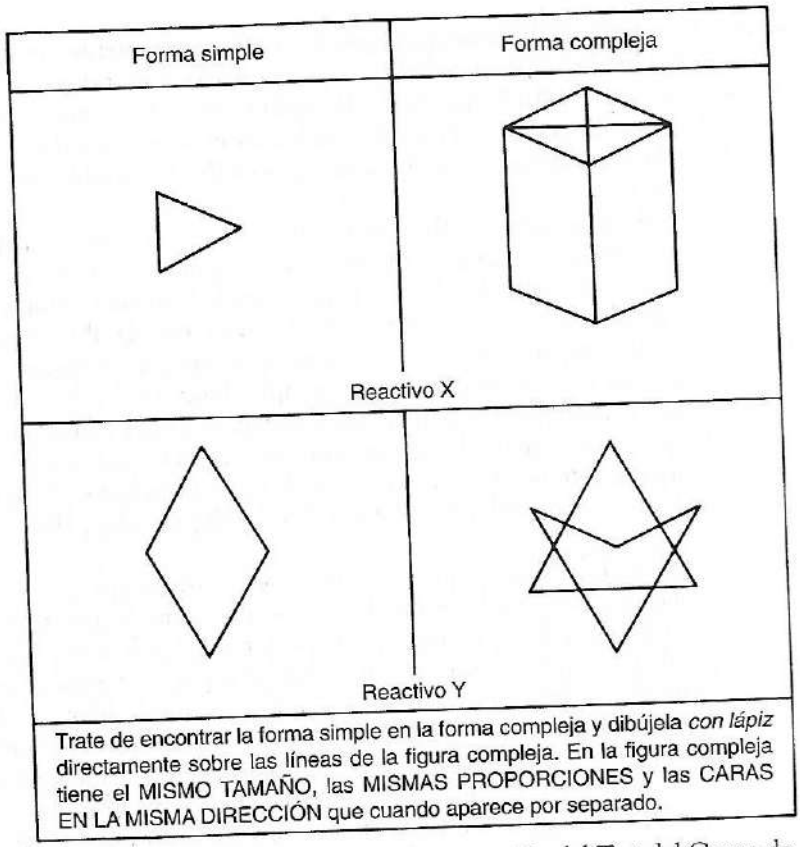


Figura 16.2. Reactivos de demostración del Test del Grupo de Figuras Encajadas.

(Copyright © 1972 por Consulting Psychologist Press. Reproducido con autorización.)

mas para niños y para preescolares las pruebas se aplican individualmente y las figuras complejas son objetos familiares; la forma original para adultos también es de aplicación individual.

Pese a la abundante investigación sobre el tema (véase, por ejemplo, Bertini, Pizzamiglio y Wapner, 1985), es mucho lo que debe hacerse para aclarar las inconsistencias entre algunos estudios de la independencia de campo y sus relaciones con diversas variables conductuales. Un problema tiene que ver con la nomenclatura, lo que a su vez refleja las diferencias en la forma en que se considera a la independencia del campo. Sobre la base de su correlación con las medidas de g, algunos investigadores han llegado a la conclusión de que el constructo de la independencia del campo es, después de todo, una habilidad cognoscitiva; otros la clasifican como un "control" cognoscitivo y la colocan en una posición intermedia entre las habilidades y los estilos cognoscitivos (McKenna, 1984; Jonassen y Grabowski, 1993). La falta de uniformidad entre los estudios en lo que respecta a las características de los participantes y la metodología es

otro obstáculo para la generalización; por ejemplo, los descubrimientos proponen que mientras los resultados de las pruebas de lápiz y papel, como el Test de Grupo de Figuras Encajadas, están muy influidos por la inteligencia general, las calificaciones en los instrumentos de ejecución, como el Portátil de la Vara y el Marco, pueden considerarse más un reflejo de la independencia de campo que estilo de variable libre de valor (Arthur y Day, 1991).

A pesar de los problemas conceptuales y metodológicos que su estudio puede presentar, los estilos cognoscitivos son de gran interés teórico y práctico.[3] Como se encuentran en la intersección entre las habilidades y la personalidad, pueden influir y modular la conducta tanto en el campo afectivo como en el intelectual, de ahí que adquieran gran importancia las interrogantes acerca de su naturaleza y sus características, así como el grado en que son fijos o flexibles (véase, por ejemplo, Niaz, 1987). Más aún, en los entornos en los que el aprendizaje es la meta principal, cada vez es mayor el reconocimiento de que la eficiencia tanto en la instrucción como en la evaluación pueden mejorar considerablemente al tener en cuenta factores de estilo (Furnham, 1995; Jonassen y Grabowski, 1993; Lu y Suen, 1995; Sternberg, 1988, 1994b; Zelniker, 1989).

Tipos de personalidad. Al igual que los estilos cognoscitivos, los *tipos de personalidad* también se refieren a constructos utilizados para explicar similitudes y diferencias en los modos preferidos de pensar, percibir y comportarse del individuo. Fundamentalmente, los tipos de personalidad son categorías definidas por configuraciones de dos o más rasgos o atributos. Como explicaciones del comportamiento humano, las tipologías tienen una larga historia que puede llevarse a la antigua teoría griega de los humores corporales (sangre, bilis negra, bilis amarilla y flema) y los temperamentos asociados con ellos: sanguíneo, melancólico, colérico y flemático. Los sistemas tipológicos a menudo resultan atractivos porque ofrecen una base sencilla y de apariencia firme para comprender y quizá explicar la conducta propia y la ajena. Estos sistemas difieren mucho en el número y la clase de las dimensiones que pueden emplear para establecer los diversos tipos.

En la psicología se han generado con los años diversas tipologías; por ejemplo el sistema de Sheldon para clasificar los tipos de personalidad a partir del físico (Sheldon y Stevens, 1942/1970) se basaba principalmente en especulaciones y se descartó por carecer de evidencias que lo corroboraran. Otros sistemas, como los códigos tipo MMPI que estudiamos en el capítulo 13, surgieron de observaciones empíricas, continuaron evolucionando y aún están en uso (Graham, 1993). Casi todos los sistemas tipológicos hacen distinciones en áreas separadas (como las ocupaciones o los temperamentos), y son muy pocos los que abarcan más de un dominio. Entre éstos sobresalen los esfuerzos de Alan Miller por sintetizar las tipologías existentes entre las dimensiones cognoscitivas, afectivas y motivacionales. Los vínculos conceptuales entre sistemas de Miller (1991a, 1991b) se fundamentan en la revisión y el análisis de una gran cantidad de tipologías en las tres áreas y pueden demostrar ser de enorme valor heurístico.

[3] Debe advertirse que los estilos cognoscitivos difieren en su complejidad. Algunos atañen a diferencias relativamente simples; por ejemplo, el estilo reflexivo contra el impulsivo, o el tempo cognoscitivo, se define sólo por la rapidez con que los individuos responden a las situaciones problemáticas, especialmente las que requieren un alto grado de incertidumbre (Kagan, 1965; S. B. Messer, 1976). Otros, como los estilos de pensamiento de Sternberg (1988, 1994b), abarcan patrones más complejos de predisposición.

Indicador de Tipos de Myers-Briggs. Jung (1921/1971) estableció una de las clasificaciones tipológicas más duraderas que sirvió de base para el Indicador de Tipos de Myers-Briggs (*Myers-Briggs Type Indicator*, MBTI; Myers, 1962; Myers y McCaulley, 1985), un instrumento muy utilizado para la evaluación de la personalidad de individuos normales.[4] El MBTI emplea la conocida dicotomía de Jung de las actitudes extravertidas y las introvertidas (E e I), su clasificación de las formas opuestas de percibir (sensación o intuición, S o I) y de las distintas maneras de acercarse al juicio (pensamiento o sentimiento, P o S); incluye además una polaridad de preferencias en la orientación al mundo externo (juicio o percepción, J o P), que no era explícita en el trabajo de Jung. Las calificaciones obtenidas en estas cuatro dimensiones, supuestamente independientes, arrojan 16 posibles "fórmulas de tipo" en las que las combinaciones de las letras indican la dirección preferida en cada una de las cuatro dimensiones; por ejemplo, la combinación "IIPP" representa al tipo "introvertido, intuitivo, pensante y perceptivo".[5] Cada dirección preferida tiene también una calificación numérica que indica la fuerza de la preferencia y se calcula a partir de la diferencia entre los puntos de los polos de cada dimensión.

A diferencia de lo que sucede en casi todos los inventarios de personalidad, el propósito de los resultados del MBTI es que sea el individuo quien haga uso de ellos, por lo que se presentan sin hacer juicios. Dos de las premisas más importantes en la interpretación de los resultados del MBTI son (a) que todos los tipos son valiosos y necesarios y que tienen puntos fuertes y puntos vulnerables especiales, y (b) que los individuos muestran mayor habilidad en las funciones, procesos y actitudes que prefieren. Estas características han contribuido a la popularidad del MBTI y a su aplicación para diversos propósitos, incluyendo el asesoramiento para la elección de carrera, la consejería y la selección y desarrollo de equipos. Los problemas psicométricos del MBTI han sido bien documentados (para revisiones, véase DeVito, 1985; J. S. Wiggins, 1989) y en particular se objetan las dicotomías forzadas del instrumento mediante las cuales se asignan las mismas letras al tipo de perfil sobre la base de diferencias en las calificaciones que pueden variar considerablemente en magnitud. No obstante, la investigación sobre el instrumento continúa, y mientras algunos estudios abordan la validez del MBTI, otros exploran formas alternativas de calificarlo (véase, por ejemplo, Davis, Grove y Knowles, 1990; Girelli y Stake, 1993; Harvey y Murry, 1994); además, se han formulado variaciones del Myers-Briggs y adaptaciones de sus constructos para diversos usos. Una variante digna de atención es el recientemente desarrollado Cuestionario de Estilos de los Estudiantes (*Student Styles Questionnaire*, Oakland, Glutting y Horton, 1996) que pretende evaluar los estilos de aprendizaje, de trabajar y de relacionarse de estudiantes de entre ocho y 17 años.

A pesar de la popularidad del Myers-Briggs y de otras tipologías, históricamente casi todos los psicólogos que han estudiado la personalidad la han concebido como una

[4] En el capítulo 13 mencionamos el MBTI en relación con el uso de la técnica de elección forzada en los inventarios autodescriptivos de personalidad.

[5] De acuerdo con el manual del MBTI, una persona así debería, entre otras cosas, (a) ser tranquila y reservada, (b) disfrutar al resolver problemas mediante la lógica y el análisis, (c) interesarse principalmente en las ideas y (d) tender a mostrar intereses bien definidos (Myers y McCaulley, 1985).

entidad psíquica formada por componentes o rasgos en los que la gente difiere cuantitativamente, más que como una taxonomía basada en diferencias cualitativas. Más aún, en la psicología científica se ha dado una renuencia general a utilizar los tipos como conceptos explicativos que surge de tres fuentes principales: (a) el énfasis que muchos teóricos de la personalidad han puesto en la necesidad de comprender y valorar la unicidad de los individuos, (b) el enlace implícito entre tipos y estereotipos y los riesgos inherentes asociados con los últimos, y (c) la escasez de métodos cuantitativos adecuados para identificar y analizar los datos de las categorías.

Sin embargo, en una serie de desarrollos recientes algunas de las objeciones tradicionales a los constructos tipológicos parecen quedar en tela de juicio; por ejemplo, Paul Meehl cree que ya es tiempo de explorar empíricamente la posible existencia de clases no arbitrarias que ocurren de manera natural, en el rango no patológico de las diferencias individuales en la personalidad; en colaboración con sus asociados, propuso nuevos métodos taxonométricos que pueden utilizarse para identificar tales clases (Meehl, 1992, 1995; Meehl y Golden, 1982; Meehl y Yonce, 1994).[6] De igual modo, después de revisar diversos modelos geométricos de rasgos de personalidad que van de los unidimensionales bipolares a los acentos multidimensionales, Dahlstrom (1995) llegó a la conclusión de que dichos modelos son inadecuados para las tareas de la investigación y la organización de los constructos de personalidad. Según él, tanto la naturaleza como el funcionamiento configurativos de la estructura de la personalidad requieren "un esquema de casillero para resumir de manera creíble y exacta la forma en que diferimos y en que nos parecemos unos a otros" (p. 14). Por último, hace poco se cuestionó incluso la noción tan arraigada de que los estereotipos suelen ser inexactos y destructivos (Lee, Jussim y McCauley, 1995).

Éstos y otros adelantos pronostican un renovado interés y una mejor aproximación al uso de categorías multidimensionales, como los tipos y estilos, como parte del caudal conceptual de que se dispone para estudiar y explicar las diferencias individuales en la conducta. Con todo, e independientemente de qué tan útiles o populares puedan ser estos constructos, es indudable que coexistirán con las aproximaciones (sin reemplazarlas) que destacan los aspectos únicos de los individuos y de su comportamiento. En la evaluación de la gente, es necesario estar siempre al tanto de los peligros de tratar de materializar los tipos y usarlos de manera simplista o como conceptos explicativos.

PRUEBAS SITUACIONALES

Aunque el término "prueba situacional" se popularizó durante y después de la Segunda Guerra Mundial, los instrumentos que se ajustan a esta descripción se desarrollaron antes de esa época. La prueba situacional tiene lugar cuando el examinado se encuentra en una situación que se parece o simula una situación de criterio "de la vida real", lo que las asemeja a las técnicas de muestra de trabajo utilizadas en la construcción de las pruebas de rendimiento ocupacional y de las que se basan en el desempeño del aprovechamiento educativo (estudiamos las dos en el capítulo 17). Sin embargo, en

[6] Aunque hay pocas evidencias relativas a la posibilidad de transportarlos de una situación a otra, los subgrupos generados mediante los análisis de agrupamientos de los datos biológicos —que veremos más tarde en este capítulo— también ofrecen posibilidades interesantes desde el punto de vista de la clasificación tipológica.

los instrumentos actuales la conducta criterio que se muestrea suele ser más variada y compleja; además, la ejecución se evalúa principalmente en términos de variables emocionales, interpersonales, actitudinales y otras variables de personalidad, más que en términos de habilidades y de conocimiento.

Pruebas de la investigación sobre la educación del carácter. Entre las primeras pruebas situacionales (aunque no se conocían con ese nombre) se encuentran las elaboradas por Hartshorne, May y sus colaboradores (1928, 1929, 1930) para la Investigación de la Educación del Carácter (*Character Education Inquiry*, CEI). Aunque estos instrumentos se diseñaron sobre todo como herramientas de investigación para un ambicioso proyecto sobre la naturaleza y el desarrollo del carácter en los niños, las técnicas han sido adaptadas a otros propósitos de evaluación.

En general, las técnicas CEI empleaban situaciones naturales y familiares de la rutina escolar del niño; las pruebas se aplicaban en forma de exámenes escolares regulares, como parte de las tareas asignadas, en el desenvolvimiento de concursos atléticos o como juegos; además, los niños no se enteraban de que los estaban examinando, excepto en la medida en que en el procedimiento pudieran incluirse exámenes escolares rutinarios. Al mismo tiempo, todos los instrumentos de Hartshorne-May representaban herramientas cuidadosamente estandarizadas que producían puntuaciones cuantitativas objetivas.

Las pruebas CEI se diseñaron de manera novedosa e ingeniosa para medir características conductuales como la honestidad, el autocontrol y el altruismo. Casi todas se interesaban por la honestidad e incluían situaciones en las que se hacía creer a los niños que podían copiar sin ser descubiertos; por ejemplo, en el rompecabezas de círculos las instrucciones dadas al niño le pedían que hiciera una marca en cada uno de 10 pequeños círculos, dispuestos irregularmente, mientras mantenía los ojos cerrados. Las pruebas en condiciones de control que impedían las miradas furtivas indicaban que era muy improbable obtener una puntuación de más de 13 marcas correctas en un total de tres ensayos, por lo que una puntuación por encima de 13 se registraba como evidencia de que había hecho trampa.

Casi todas las pruebas CEI demostraron tener buena capacidad discriminativa (lo que dio lugar a una amplia gama de diferencias individuales en las puntuaciones) y también su confiabilidad resultó bastante satisfactoria; sin embargo, las respuestas de los niños mostraban una considerable especificidad situacional. Así, las correlaciones de las diferentes pruebas en cada categoría (por ejemplo, pruebas de honestidad o de persistencia) demostraban ser muy bajas. Esta especificidad puede entenderse cuando consideramos la operación de los intereses, valores y motivos del niño en las distintas situaciones. Por ejemplo, al niño motivado por destacar en el trabajo escolar no necesariamente le preocupa el rendimiento en concursos atléticos o en juegos de fiestas. Estas diferencias motivacionales, a su vez, deberían verse reflejadas en la conducta del niño en las pruebas de honestidad aplicadas en diferentes contextos. Tanto en sus hallazgos como en sus interpretaciones, los investigadores de la CEI anticiparon por cuatro décadas la importancia de la especificidad situacional. Comoquiera que sea, los nuevos análisis de los datos de la CEI indican que un modelo más apropiado incluye tanto un factor general de honestidad como un componente situacional, y hay incluso

ciertas evidencias a favor de la existencia de un factor general del carácter moral (Burton, 1963; Rushton, 1984).[7]

Pruebas situacionales en los centros de evaluación y técnicas de representación de papeles. Las pruebas situacionales constituyen una parte importante del programa del centro de evaluación introducido en los Estados Unidos por la Oficina de Servicios Estratégicos (*Office of Strategic Services*, OSS) durante la Segunda Guerra Mundial. En esencia, la técnica del centro de evaluación consiste en un periodo de "estancia" durante el cual a los candidatos se los observa y examina. Éste era el procedimiento principal en la selección del personal militar para la asignación de tareas cruciales en el extranjero (Murray y MacKinnon, 1946; OSS Assessment Staff, 1948). Después se utilizaron procedimientos similares en el Instituto para la Evaluación de la Personalidad y la Investigación en la Universidad de California y en diversos proyectos de investigación a gran escala para personal militar y civil.[8] Y más tarde, los grandes consorcios industriales adoptaron la técnica del centro de evaluación y algunas pruebas situacionales para la evaluación de sus ejecutivos de nivel superior.

La OSS elaboró la prueba de *estrés situacional* para muestrear la conducta del individuo en condiciones de estrés, frustrantes o de trastornos emocionales. En una de ellas, se asigna al examinado una tarea que debe realizar con dos "ayudantes" que no sólo no cooperan, sino que además lo obstaculizan. Otra prueba situacional emplea un *grupo sin líder* como instrumento para valorar características como el trabajo en equipo, la habilidad, la iniciativa y el liderazgo. En estas pruebas, la tarea asignada requiere los esfuerzos cooperativos de un grupo de examinados, ninguno de los cuales es designado como líder o recibe responsabilidades específicas. Algunos ejemplos del programa OSS incluyen la Situación Arroyo, que requiere la transferencia de personal y equipo por un arroyo a la máxima velocidad y con la mayor seguridad, y la Situación Muro, en la que la gente y los materiales tienen que ser transportados a través de una pared doble separada por un abismo imaginario.

La Discusión Grupal sin Líder (DGL) es una variante de la técnica anterior que requiere un mínimo de tiempo y de equipo, por lo que se ha usado mucho en la selección de grupos como oficiales militares, supervisores y administradores de servicio civil, ejecutivos industriales y gerentes expertos, vendedores entrenados, maestros y trabajadores sociales. En esencia, consiste en asignar al grupo un tema que debe discutirse durante un periodo específico. Los evaluadores observan y califican la ejecución de cada participante pero no intervienen en la discusión. Aunque a menudo se usa en condiciones informales y no estandarizadas, la técnica ha sido sometida a una considerable investigación cuyos resultados indican que, sobre todo cuando los calificadores están adecuadamente capacitados, puede ser un instrumento eficaz para predecir la ejecución en trabajos que requieren comunicación verbal, resolución de problemas verbales y aceptación de los pares (Bass, 1954; Greenwood y McNamara, 1967; Guilford, 1959; Thornton y Zorich, 1980).

---

[7] El lector puede encontrar en Vitz (1990) un análisis interesante del papel que han desempeñado los estudios de Hartshorne y May en el debate sobre la enseñanza de virtudes y el desarrollo del carácter.

[8] En J. S. Wiggins (1973/1988 capítulo 11) puede encontrarse una explicación de esos importantes estudios de evaluación, empezando con el programa OSS.

Algunas pruebas situacionales emplean la técnica de representación de papeles (*role-playing*), una forma de improvisación para producir la conducta de interés. De hecho, los grupos de discusión sin líder —que acabamos de citar— así como algunas otras pruebas situacionales pueden considerarse variantes o elaboraciones de la representación de papeles que fue una de las técnicas utilizadas en el programa de evaluación OSS, si bien sus orígenes son anteriores y tiene aplicaciones más amplias. En McReynols y DeVoge (1978) puede encontrarse una revisión minuciosa de la historia, lógica y variedades de la técnica de la improvisación, en las que se dan instrucciones explícitas al individuo de que represente un papel, sea de manera abierta (con o sin la presencia de otras personas) o mediante el informe verbal de lo que haría o diría. La situación puede presentarse de manera realista o bien por medio de cintas de audio, de video o una descripción impresa.[9]

La técnica de improvisación sigue gozando de gran popularidad, aunque gran parte de su uso es informal y está adaptado a ambientes específicos y condiciones locales. Una aplicación importante se encuentra en la evaluación de personal, especialmente cuando el trato con la gente es una función importante del empleo (Stricker, 1982; Stricker y Rock, 1990). Un ejemplo especial tiene que ver con la evaluación de la eficacia de los consejeros. En esta situación, el consejero prospecto es observado (o filmado) mientras conduce una sesión con un "cliente asesorado", es decir, un miembro del equipo o un compañero que desempeña el papel del cliente y que presenta un problema estandarizado y seleccionado de antemano (véase, por ejemplo, Connor, 1994, pp. 72–75; Kelz, 1966; Neufeldt, Iversen y Juntunen, 1995; A. Williams, 1995). Cuando se filma la sesión, además de la calificación que recibe de asesores y pares, el propio sujeto puede supervisar y evaluar su desempeño.[10]

Las técnicas del centro de evaluación han sido utilizadas en los campos en los que los criterios de selección son complejos, por ejemplo, el trabajo en las fuerzas policiacas (J. L. Coleman, 1987; Moore y Unsinger, 1987), y reiteradamente han demostrado que pueden predecir con éxito diversos criterios (véase, por ejemplo, Coulton y Feild, 1995; Howard y Bray, 1988; Ritchie, 1994; Tziner, Ronen y Hacohen, 1993). Incluso se han hecho esfuerzos para adaptarlas a la evaluación de solicitantes de empleo sordos (Berkay, 1993). Debido a la amplitud de sus aplicaciones, no podemos generalizar acerca de la validez de *la* técnica del centro de evaluación; los resultados difieren de un centro a otro dependiendo de los procedimientos empleados, la naturaleza de los criterios y la calificación de los asesores. En general, los coeficientes de validez tienden a ser más altos cuando la metodología es la adecuada (como en los estudios que usan varios instrumentos, incluyendo la evaluación de los pares) y se concentra en las dimensiones conductuales relevantes y observables directamente (Gaugler, Rosenthal, Thornton y Bentson, 1987; Shore, Shore y Thornton, 1992; Thornton y Byham, 1982). A pesar de la vasta investigación, aún quedan sin responder algunas interrogantes acerca de las

[9] Los multimedios y las tecnologías interactivas de las computadoras hicieron posible el surgimiento de formas completamente nuevas de presentar situaciones de estímulo realistas y opciones de respuesta para los examinados. Véase Drasgow, Olson-Buchanan y Moberg (1996), y Olson-Buchanan, Drasgow, Moberg, Mead y Keenan (1996) para una revisión de las posibilidades y problemas en el desarrollo de pruebas en multimedios y para una descripción de un instrumento en video interactivo que permite medir las habilidades para la resolución de conflictos.

[10] En la psicología clínica, la representación de papeles ha pasado por un desarrollo extenso y sistemático en diversas orientaciones teóricas diferentes, particularmente en los programas de modificación de conducta, la terapia familiar y la consejería matrimonial.

técnicas del centro de evaluación, de las cuales las más inquietantes son las que tienen que ver con la incapacidad de varias investigaciones para demostrar la validez convergente entre diferentes métodos para evaluar dimensiones particulares del desempeño.[11]

AUTOCONCEPTO Y CONSTRUCTOS PERSONALES

Los años ochenta y noventa han presenciado el resurgimiento del interés en el autoconcepto y constructos relacionados (Byrne, 1996; Harter, 1990; Hattie, 1992; Markus y Wurf, 1987; Oosterwegel y Oppenheimer, 1993; Wylie, 1989).[12] Muchas de las aproximaciones actuales a la evaluación de la personalidad prestan mucha atención a la forma en que el individuo se ve y se acepta a sí mismo y a los demás. En este sentido, las técnicas reflejan la influencia de la psicología fenomenológica, que concede gran importancia a la manera en que el individuo *percibe* los acontecimientos. Tenemos entonces que la descripción que la persona hace de ella misma adquiere una importancia por derecho propio en lugar de considerarse como un sustituto de otras observaciones conductuales.

Los procedimientos considerados en esta sección tienen su foco principal en la forma en que los individuos se perciben a sí mismos y a los demás. Aunque algunos de los instrumentos descritos se publicaron como pruebas, muchos otros se desarrollaron para emplearse en proyectos específicos de investigación y pueden encontrarse en los informes publicados. Algunos son de interés principalmente por su asociación con determinadas teorías de la personalidad o con un área activa de investigación, mientras que otros constituyen técnicas de gran aplicación que se utilizan en el estudio de diversos problemas.

Test de Completamiento de Oraciones de la Universidad de Washington. En esencia, puede argumentarse que las pruebas de autoconcepto no difieren de los inventarios autodescriptivos que revisamos en el capítulo 13. Aunque esto es cierto, sería más exacto decir que estos inventarios son medidas del autoconcepto. La interpretación de las respuestas al inventario en términos del autoconcepto establece la base de la aproximación teórica al desarrollo de la personalidad propuesta por Loevinger (1966a, 1966b, 1976, 1987, 1993; Loevinger y Ossorio, 1958). Al unir muchos descubrimientos diferentes de su propia investigación y del trabajo de otros, Loevinger propuso un rasgo de personalidad que definió como la capacidad para conceptuarse o para "tomar distancia" de uno mismo y de los propios impulsos. De acuerdo con el investigador, las manifestaciones de este rasgo en los inventarios de personalidad se han descrito como apariencias, posiciones defensivas ante la prueba, predisposición de respuesta, deseabilidad social, aquiescencia y estilo personal. Al igual que muchos otros psicólogos, Loevinger considera que estas actitudes ante la situación de prueba no constituyen errores instrumentales que deben ser eliminados, sino la fuente más importante de varianza válida en los inventarios de personalidad.

---

[11] El lector puede encontrar en Landy, Shankster y Kohler (1994, pp. 277–278), y en Schmidt, Ones y Hunter (1992, pp. 635–637) revisiones breves pero informativas de la investigación sobre el centro de evaluación.

[12] En Levin (1992) puede hallarse un repaso general de las aproximaciones filosóficas y psicológicas a los procesos del yo y otros procesos relacionados, del siglo XVII al XX.

Tomando como base los datos de muchas fuentes, Loevinger propuso que la habilidad para formar un autoconcepto aumenta con la edad, la inteligencia, la educación y el nivel socioeconómico. En el punto más bajo, ilustrado por el infante, el individuo es incapaz de formar un autoconcepto. A medida que aumenta su habilidad, el niño va formando un concepto estereotipado, convencional y socialmente aceptable de sí mismo, una etapa que Loevinger considera característica de la adolescencia. Al aumentar su madurez, el individuo progresa del concepto estereotipado a un autoconcepto diferenciado y realista, momento en que el individuo está totalmente consciente de sus peculiaridades y se acepta por lo que es.

Loevinger y sus colaboradores pretenden medir este rasgo del autoconcepto, conocido como desarrollo o nivel del yo, en el Test de Completamiento de Oraciones de la Universidad de Washington (*Washington University Sentence Completion Test*, WUSCT, Loevinger, 1985, 1987; Loevinger y Wessler, 1970; Loevinger, Wessler y Redmore, 1970). El marco teórico de los autores plantea nueve niveles de desarrollo del yo: Presocial, Impulsivo, Autoprotector, Conformista, Consciente de sí mismo, Escrupuloso, Individualista, Autónomo e Integrado, cada uno de los cuales (salvo el primero, que precede a la aparición de las habilidades verbales) pueden evaluarse con el WUSCT. A cada completamiento de las oraciones se les asigna una calificación de nivel del yo y a partir de estos resultados se calcula una calificación compuesta de toda la prueba. El WUSCT se basó en la investigación realizada con mujeres y chicas y luego se adaptó para hombres y jóvenes adolescentes. En nuestros días ha sido revisado y ahora proporciona formas comparables para ambos géneros (Loevinger, 1985; Novy, 1992) y un nuevo manual de calificación presenta datos para mujeres y para hombres (Hy y Loevinger, 1996).

La investigación subsecuente con el WUSTC ha arrojado evidencias que apoyan su confiabilidad (Novy y Francis, 1992; Weiss, Zillberg y Genevro, 1989) y su validez como medida del constructo del desarrollo del yo entre varias muestras (Bushe y Gibbs, 1990; Novy, Gaa, Frankiewicz, Liberman y Amerikaner, 1992; Westenberg y Block, 1993). El instrumento ha resultado particularmente exitoso en el estudio de las diferencias de género en el desarrollo del yo (Cohn, 1991). Uno de los pocos aspectos problemáticos del WUSCT, que comparte con otras pruebas de completamiento de oraciones y técnicas que permiten respuestas verbales abiertas, es que sus calificaciones tienden a correlacionar y, en algunos casos, a verse afectadas por la fluidez verbal y el nivel del vocabulario (véase, por ejemplo, Vaillant y McCullough, 1987; Westenberg y Block, 1993). Aunque es necesario tomar en cuenta esta posibilidad al diseñar la investigación y el empleo del WUSTC, está en concordancia con sus bases teóricas, por lo que no necesariamente invalida los resultados.

Inventarios de autoestima y medidas relacionadas. En la siempre creciente investigación, el autoconcepto se mezcla con otros constructos relacionados como la autoestima (Baumeister, 1993; Bednar y Peterson, 1995) y la autoeficacia percibida (Bandura, 1982, 1995; Maddux, 1995; Schwarzer, 1992). Un punto fuerte de esta línea de investigación tiene que ver con el efecto de la evaluación que el individuo hace de su propia ejecución; de hecho, suele describirse la autoestima como el componente evaluativo del autoconcepto que a largo plazo puede influir en el desarrollo de los rasgos cognoscitivos y afectivos. En particular, existe un amplio acuerdo respecto a la

función crucial que juega la autoestima en la determinación de variables psicológicas tan importantes como la capacidad de afrontamiento y la sensación de bienestar.

El constructo de la autoestima es, a primera vista, engañosamente simple. A menudo se supone que es una actitud general de autoevaluación, que fluctúa entre lo muy positivo y lo muy negativo, que es estable y de naturaleza por completo subjetiva. Tales suposiciones han sido tradicionalmente la base de la medición de la autoestima con propósitos de investigación y de intervención, al grado que en su extensa revisión de las medidas de autoestima y autoconcepto Blascovich y Tomaka (1991) señalaron a la Escala de Autoestima (*Self-Esteem Scale*, SES) de Rosenberg (1965) —una escala autodescriptiva de 10 reactivos con validez de facie— como la prueba de este tipo más utilizada.

Las medidas simples y evidentes de la autoestima global, como la obtenida con la Escala de Autoestima, parecen ser relativamente estables en el tiempo y para ciertos propósitos pueden ser útiles como medida de la forma general en que el individuo se ve a sí mismo. No obstante, muchos investigadores han encontrado que la relación entre conducta y autoconcepto puede demostrarse mejor si éste se considera y evalúa como un constructo jerárquico y multidimensional (Fleming, Courtney, 1984; Marsh, Byrne y Shavelson, 1992; Marsh y Shavelson, 1985; Shavelson y Bolus, 1982; Shavelson, Hubner y Stanton, 1976; Uguroglu y Walberg, 1979). En algunas circunstancias, el uso de una sola medida global de la autoestima puede arrojar resultados inconsistentes o no revelar correlaciones significativas con otras variables, que tal vez aparezcan como consistentes y significativas con un constructo más definido, como el autoconcepto académico. Esto es particularmente cierto en el caso de aquellas personas para las cuales el aprovechamiento académico puede no ser tan importante en su sistema personal de valores; en estos casos, es posible que un autoconcepto global elevado, que tiende a reflejar el sistema de valores del examinado, no muestre una correlación significativa con el aprovechamiento académico o el funcionamiento intelectual. Más aún, la investigación indica que no siempre hay una correlación entre las medidas de la "autoestima experimentada", que se basan en autodescripciones, y las medidas de "autoestima presentada", que atañen a las autodescripciones de otros (Demo, 1985).

Como resultado de estos descubrimientos, en los años recientes la investigación se ha alejado de la conceptuación unidimensional de la autoestima para dirigirse al estudio de sus facetas específicas, y ha proliferado rápidamente. Las investigaciones de las relaciones entre el autoconcepto y el aprovechamiento académicos en niños y adolescentes constituyen un área en particular activa en la que los resultados de los estudios que emplean metodologías complejas (como el modelamiento de ecuaciones estructurales y los diseños longitudinales que permiten el análisis de la dirección de las relaciones causales) respaldan la idea de que el autoconcepto académico está vinculado a los campos específicos de estudio; además, se ha demostrado que este concepto correlaciona, predice e influye en el aprovechamiento académico subsecuente (Fortier, Vallerand y Guay, 1995; House, 1995; Lyon y MacDonald, 1990; H. W. Marsh, 1990a, 1990b).

Los instrumentos más recientes para evaluar el autoconcepto hacen uso de la extensa bibliografía teórica y empírica acumulada en el área; por ejemplo, la Escala de Autoconcepto del Estudiante (*Student Self-Concept Scale*, SSCS; Gresham, Elliott y Evans-Fernández, 1993) es una prueba que parte de la teoría de Bandura (1982,

1986) de la autoeficacia (aunque también toma elementos de otras teorías e investigaciones) para evaluar tres áreas principales del autoconcepto: académica, social y de autoimagen. En cada una de estas áreas, los individuos indican qué tanta confianza sienten en su capacidad para hacer lo que afirman los reactivos, qué tan importantes son estos reactivos para ellos y qué tanta confianza tienen de poseer ciertos atributos o hacer determinadas cosas que les permitan obtener ciertos resultados. La SSCS produce calificaciones compuestas referidas a normas, y subescalas para estudiantes de ambos géneros de la escuela elemental y de secundaria.

El Inventario de Autoconcepto Personal y Académico (*Personal and Academic Self-Concept Inventory*, PASCI, Fleming y Whalen, 1990), es una prueba similar desarrollada con estudiantes de secundaria y universitarios que pretende investigar y hacer operativo el modelo jerárquico multifacético del autoconcepto de Shavelson y sus colegas. El PASCI se encuentra en la cuarta revisión y expansión de una escala experimental que se originó en la década de los cincuenta para evaluar los sentimientos de inadecuación.[13] La versión actual consiste en una escala global de la autoestima y seis escalas adicionales de facetas, de las cuales dos tratan con los aspectos sociales del autoconcepto (aceptación social y ansiedad social), dos atañen a los aspectos físicos (apariencia y capacidad física) y las dos restantes son académicas (habilidad para las matemáticas y habilidad verbal). Por supuesto, tanto la SSCS como el PASCI están sujetos a los sesgos de respuesta inherentes a todas las medidas autodescriptivas (que revisamos en el capítulo 13). Ambos instrumentos necesitan además documentar su eficacia en los contextos para los que fueron desarrollados; sin embargo, desde el punto de vista de la diferenciación del constructo y de la especificación de contenido, son representativos de los avances realizados en la conceptuación y medición de la autoestima.

La lista de verificación de adjetivos. Se han formulado algunas técnicas de amplia orientación para la evaluación del autoconcepto, tal es el caso de la Lista de Verificación de Adjetivos (*Adjective Check List*, ACL), un instrumento de gran aplicación. Preparada originalmente para emplearse en el programa de investigación del Instituto para la Evaluación e Investigación de la Personalidad (*Institute for Personality Assessment and Research*, IPAR), presenta una lista alfabética de 300 adjetivos, que van de "atolondrado" a "zoquete" (Gough, 1960; Gough y Heilbrun, 1983), en la que los examinados marcan todos los adjetivos que consideran que los describen.

En su forma actual, la ACL puede ofrecer calificaciones para 37 escalas, cuatro de las cuales evalúan las predisposiciones de respuesta. Originalmente se preparó un agrupamiento principal de escalas sobre una lógica o base del contenido al asignar los adjetivos a cada una de las 15 necesidades de Murray cubiertas por el EPPS (véase el capítulo 13) y se desarrolló un conjunto adicional de nueve "escalas temáticas" para medir los diversos rasgos que se consideran importantes en la conducta interpersonal. Los dos agrupamientos restantes se diseñaron para ajustarse a las teorías especializadas de la personalidad, a saber, la teoría del análisis transaccional de Berne (1961, 1966) y

[13] Blascovich y Tomaka (1991) proporcionan un resumen de la evolución de este instrumento, así como revisiones de muchas otras pruebas de autoestima. El modelo de Shavelson ha sido el de mayor influencia en el campo y también se ha utilizado en la conceptuación de la SSCS. Para mayor información sobre el modelo, véase Marsh *et al.* (1992), Marsh y Shavelson (1985), Shavelson y Bolus (1982) y Shavelson *et al.* (1976).

la teoría de la creatividad y la inteligencia de Welsh (1975b). Para las 37 escalas, la fuente principal de datos de validación empírica, utilizada en algún momento del desarrollo de cada escala, fue la observación directa de los participantes en los programas del centro de evaluación en el IPAR y las valoraciones resultantes de los rasgos. Sobre la base de estas evaluaciones IPAR así como en otras investigaciones complementarias, el manual proporciona descripciones de la personalidad de los individuos que obtienen calificaciones altas y bajas en cada escala.

Como instrumento de investigación, la ACL se ha aplicado a una sorprendente variedad de problemas derivados de áreas: psicopatología, elección ocupacional, creatividad, conducta política y económica e incluso las reacciones de los pacientes a la ortodoncia y a los lentes de contacto. También se ha empleado en la valoración de personajes históricos a partir de sus biografías y de sus obras publicadas (Welsh, 1975a) y para caracterizar objetos inanimados como ciudades y automóviles. Más recientemente, la ACL se ha utilizado para estudiar los cambios del desarrollo en mujeres de edad madura (Helson y Wink, 1992; Wink y Helson, 1993; York y John, 1992) y en las investigaciones sobre el narcisismo (Wink, 1991, 1992). El lector interesado puede encontrar una revisión del tema en Teeter (1985) y en Zarske (1985).[14]

Clasificación Q. Stephenson (1953) estableció la clasificación Q (*Q Sort*), otra técnica para la investigación del autoconcepto que pone en práctica una aproximación a la investigación conocida como metodología Q (véase, por ejemplo, Kerlinger, 1986, capítulo 32; McKeown y Thomas, 1988). En esta técnica, el examinado recibe un conjunto de tarjetas en las que aparecen afirmaciones o nombres de rasgos que debe clasificar en pilas según considere que son "muy características" o "muy poco características" de sí mismo. Aunque los reactivos pueden provenir de una lista estándar, es más común que se redacten para ajustarlos al caso individual. Para asegurar la distribución uniforme de las calificaciones se utiliza una distribución "normal forzada" en la que se pide al examinado que coloque un número específico de tarjetas en cada pila. La distribución puede prepararse, por ejemplo, para muestras de reactivos de cualquier tamaño haciendo referencia a la tabla de la curva normal. Debe advertirse que, como en la técnica de elección forzada que vimos en el capítulo 13, la clasificación Q produce datos ipsativos más que normativos. En otras palabras, los examinados informan cuáles consideran que son sus rasgos fuertes y cuáles los débiles, pero no qué tan fuertes creen ser en comparación con otras personas o con ciertas normas externas.

Las clasificaciones Q se han utilizado para estudiar una gran variedad de problemas psicológicos (Bem y Funder, 1978; Block, 1961/1978; Kogan y Block, 1991; Ozer, 1993; Rogers y Dymond, 1954). En la investigación exhaustiva de la personalidad individual, es común que se pida al individuo que vuelva a clasificar el mismo conjunto de reactivos en diferentes marcos de referencia; por ejemplo, los reactivos pueden clasificarse de modo que se apliquen al propio individuo o a otras personas como el padre, la madre o el cónyuge, o bien que se apliquen al individuo en diferentes medios, como el trabajo, el hogar o las situaciones sociales. De igual manera, pueden obtenerse clasificaciones Q para la forma en que los individuos creen que son verda-

[14] El editor de la prueba (Gough y Heilbrun, 1980) también ha puesto en circulación una bibliografía completa de la ACL hasta 1980.

deramente (el yo real), cómo creen que otros los consideran (el yo social) y de qué manera les gustaría ser (el yo ideal). Las clasificaciones Q pueden obtenerse en las diferentes etapas de la psicoterapia para observar los cambios, un procedimiento seguido especialmente por los terapeutas centrados en el cliente. Con la terapia, el autoconcepto tiende a volverse más favorable y a parecerse más al autoconcepto ideal del individuo (Rogers y Dymond, 1954, capítulo cuatro).[15]

El diferencial semántico. Esta técnica fue desarrollada primero por Osgood y sus colaboradores (Osgood, Suci y Tannenbaum, 1957) como instrumento para investigar la psicología del significado, aunque pronto se reconocieron sus posibilidades para la evaluación de la personalidad. El diferencial semántico representa un procedimiento estandarizado y cuantificado para medir las connotaciones que cualquier concepto determinado tiene para el individuo. Cada concepto se valora en una escala gráfica de siete puntos según se relacione más con uno u otro de un par de opuestos, como se ilustra en la figura 16.3. Para cada concepto se emplea una serie de estas escalas bipolares de adjetivos, y por lo general se incluyen 15 escalas o más. Las correlaciones y los análisis factoriales del conjunto original de 50 escalas de Osgood revelaron la existencia de tres factores principales: *Evaluativo*, con cargas elevadas en escalas como bueno–malo, apreciable–despreciable y limpio–sucio; *Potencia*, encontrado en escalas como fuerte–débil, grande–pequeño y pesado–ligero; y *Actividad*, identificado en escalas como activo–pasivo, rápido–lento y agudo–torpe. Además de ser el más conspicuo, el factor evaluativo es el que explica el mayor porcentaje de la varianza total.

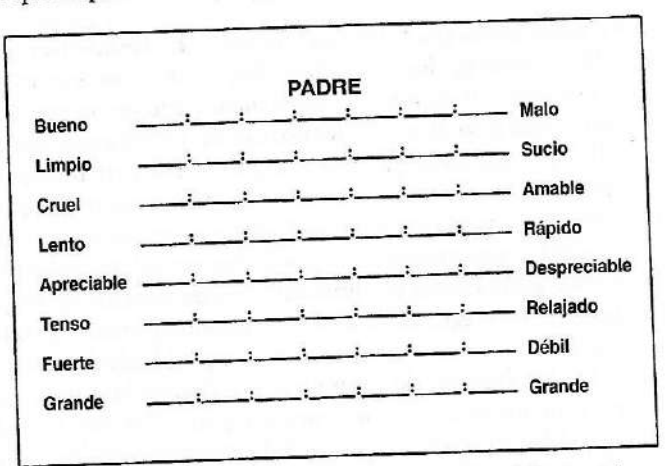


Figura 16.3. Ilustración de la técnica de diferencial semántico. Al valorar el concepto de "padre", el examinado coloca una marca en el segmento apropiado de cada escala. Por lo general se emplea un número mucho más grande de escalas.

[15] Por supuesto, todos esos procedimientos pueden seguirse junto con la Lista de Verificación de Adjetivos o con cualquiera otra de las técnicas para la evaluación del autoconcepto. Además, como veremos más tarde, pueden usarse las clasificaciones Q y las listas de verificación para informar de las evaluaciones del observador.

Las respuestas en el diferencial semántico pueden analizarse de diversas maneras. Para un tratamiento cuantitativo, a las calificaciones de cada escala se les asigna valores numéricos de uno a siete o de –3 a +3; a continuación, se mide la similitud global de dos conceptos cualquiera para un individuo o grupo en términos de su posición en todas las escalas. Las connotaciones de todos los conceptos calificados por un solo individuo pueden investigarse al calcular la "puntuación" de cada concepto en los tres factores principales que señalamos arriba. De este modo, en una escala que se extiende de –3 a +3, el concepto que el individuo tenga de, digamos, "mi hermano", puede calificar –2 en el factor evaluativo, 0.1 en potencia y 2.7 en actividad.

Los conceptos a evaluar deben elegirse de modo que se ajusten al problema que se investiga; por ejemplo, puede solicitarse a los examinados que se califiquen a sí mismos, a miembros de su familia, amigos, empleadores, maestros o figuras públicas, miembros de distintos grupos étnicos o culturales, personas que realizan diferentes ocupaciones, actividades como estudiar o hacer deporte al aire libre; ideas abstractas como confusión, aborrecimiento, enfermedad, paz o amor; nombres de productos o de marcas, y programas de radio y televisión. El diferencial semántico se ha aplicado en muchos contextos distintos, en la investigación de problemas como el diagnóstico clínico y la terapia, las elecciones vocacionales, las diferencias culturales y la reacción de los consumidores al nombre de marcas y productos (Snider y Osgood, 1969). Más aún, se sigue perfeccionando la técnica (véase, por ejemplo, Cogliser y Schriesheim, 1994), lo que se refleja en el hecho de que la bibliografía del diferencial semántico incluye más de 2 000 referencias.

Test de Repertorio de Construcción de Roles. El Test de Repertorio de Construcción de Roles (*Role Construct Repertory Test*, Rep Test) es una técnica diseñada por G. A. Kelly (1955, 1963, 1970) para fungir específicamente como auxiliar en la práctica clínica. El desarrollo del Rep Test está íntimamente relacionado con la teoría de la personalidad de Kelly, que propone que los conceptos o constructos de los que el individuo se vale para percibir los objetos o acontecimientos pueden influir en su comportamiento. En el curso de la psicoterapia, muchas veces es necesario elaborar nuevos constructos y descartar algunos antes de que pueda hacerse algún progreso.

El Rep Test se diseñó para ayudar al clínico a identificar algunos de los constructos más importantes del cliente acerca de la gente. Aunque la prueba puede aplicarse de muchas formas, incluidas las versiones colectiva e individual, de una u otra manera siempre comprende la clasificación de estímulos. Los datos obtenidos con el Rep Test permiten evaluar las relaciones entre constructos y generalmente se ordenan en una matriz o cuadrícula. Las características básicas del procedimiento pueden ilustrarse en la descripción de una de sus variantes más sencillas,[16] en la que primero se presenta al examinado la Lista de Títulos de Roles y se le pide que nombre a una persona que, según su experiencia, se ajuste a cada uno. Entre los roles clásicos se incluyen los del padre, la madre, la pareja actual, un maestro que le agrada y una persona con la que haya tratado hace poco y a la que parece disgustarle. El examinador elige luego a tres de las personas mencionadas y pregunta: "¿En qué punto *importante* coinciden dos de estas

[16] El lector puede encontrar descripciones de otras variantes del procedimiento en Bannister y Mair (1968), Beail (1985), Landfield y Epting (1987), G. J. Neimeyer (1989) y Winter (1992).

personas y difieren de la tercera?" Este procedimiento se repite con muchos otros conjuntos de tres nombres, pues algunos reaparecen en diferentes combinaciones. Una vez que se han generado los datos crudos de los constructos importantes y se han organizado en una cuadrícula, pueden analizarse sus implicaciones. Aunque el Rep Test produce una gran cantidad de datos cualitativos, para establecer qué tan importantes son para el individuo los constructos producidos, se han desarrollado procedimientos cuantitativos que van de la simple estadística descriptiva a análisis estructurales sumamente complejos. Se dispone de programas de cómputo para ayudar tanto en el análisis de contenido como en el análisis estructural de los protocolos (véase, por ejemplo, Winter, 1992).

El Rep Test se ha empleado en sus diversas variantes en una gran cantidad de investigaciones sobre problemas relacionados con la teoría de la personalidad, la cognición social, la educación y la comunicación, la psicoterapia y la evaluación. Entre los índices derivados de las clasificaciones que hacen los sujetos de las personas familiares hay uno que se conoce como *complejidad cognoscitiva*, que se basa en el número de diferentes constructos empleados por el individuo y se considera una medida de su estilo cognoscitivo. Un mayor grado de complejidad cognoscitiva significa que el individuo emplea más dimensiones y, por consecuencia, un sistema cognoscitivo más diferenciado, para organizar o representar el entorno (Bieri, 1971; Bieri *et al.*, 1966; Goldstein y Blackman, 1978a, pp. 483–487; 1978b); sin embargo, la complejidad cognoscitiva es en sí misma un constructo multifacético y su grado de correlación con otras variables depende en gran medida de la forma en que se le conceptúe y se le mida (véase, por ejemplo, Goldsmith y Nugent, 1984).

Después de un periodo de relativa calma, la década de los ochenta presenció el resurgimiento del interés en la teoría de Kelly sobre los constructos personales, y desde entonces la investigación realizada en este marco ha continuado a paso creciente (Bannister, 1985; Burr y Butt, 1992; Epting y Landfield, 1985; Fransella y Thomas, 1988; G. J. Neimeyer y R. A. Neimeyer, 1990; R. A. Neimeyer y G. J. Neimeyer, 1992). La teoría de Kelly también ha sido uno de los pilares del constructivismo.[17] Como resulta ideal para estudiar las formas únicas en las que el individuo asigna y organiza el significado, el propio Rep Test ha pasado por un uso intenso en los últimos años, como el análisis de grupos jerárquicos y el escalamiento multidimensional (véase, por ejemplo, Merluzzi, 1991; Ogilvie y Ashmore, 1991); sin embargo, son tantas las variaciones a las que ha estado sometida la técnica que no pueden hacerse generalizaciones acerca de su eficacia y sus propiedades psicométricas.

Ambiente percibido y clima social. El concepto de ambiente llega a la psicología por diversas vías.[18] Ya hemos considerado la importante función que puede desempeñar la evaluación que los observadores hagan del ambiente en el contexto de la

[17] El constructivismo es un movimiento activo más que controversial en la psicología que considera a los seres humanos como individuos activos cuyo principal propósito es dar significado a la experiencia, y que apela a puntos de vista sobre el conocimiento diferentes al positivismo (Gergen, 1985; Mahoney, 1991; G. J. Neimeyer, 1993; R. A. Neimeyer y Mahoney, 1995).

[18] En efecto, luego de tres décadas de actividad, el estudio del ambiente y la conducta es un campo multidisciplinario que florece en todo el mundo (véase, por ejemplo, Bonnes y Secchiaroli, 1995; Groat, 1995; McAndrew, 1993; Stokols, 1995; Stokols y Altman, 1987).

evaluación de poblaciones especiales (capítulo 9), así como el impacto que las variables culturales y situacionales tienen sobre la expresión de los rasgos cognoscitivos y afectivos (capítulos 9, 11, 12 y 13). Desde un punto de vista fenomenológico, la evaluación de los ambientes y climas sociales mediante el informe que hagan las personas de sus percepciones en cada medio también puede contribuir significativamente a la comprensión de los individuos y los grupos.

Se han desarrollado diversas pruebas para describir y evaluar los aspectos físicos y sociales del ambiente (el lector puede encontrar revisiones de varias de esas pruebas en Walsh y Betz, 1995, capítulo 11). Las 10 Escalas de Clima Social (*Social Climate Scales*) que preparó Rudolph H. Moos en la Universidad de Stanford (1974, 1993a, 1993b, 1993c, 1994a; Moos y Spinrad, 1984) se encuentran entre los instrumentos de este tipo más versátiles y más utilizados. Las Escalas del Clima Social pueden aplicarse en contextos como los programas de tratamiento psiquiátrico comunitarios o basados en la hospitalización, la atención en instituciones de asilo o correccionales, ambientes militares, residencias estudiantiles en las universidades, aulas en las secundarias, medios laborales y en las familias; también se cuenta con una escala más general de ambiente grupal para grupos de apoyo mutuo, sociales y orientados a las tareas. Casi todas estas escalas constan de 90 o 100 reactivos de cierto o falso en las que el individuo describe la forma en que percibe determinado ambiente. Los reactivos se prepararon primero para cubrir las dimensiones ambientales teóricamente elegidas, como la presión para comprometerse, para la autonomía o para el orden; los últimos reactivos se eligieron de modo empírico en función de su capacidad para discriminar contextos y por la consistencia interna entre subescalas. Cada una de las 10 escalas ha sido revisada y actualizada por lo menos una vez (Moos, 1994b; Moos y Moos, 1994; Trickett y Moos, 1995). Las Escalas de Clima Social pueden aplicarse en tres formas diferentes que miden (a) las percepciones actuales del medio real, (b) las percepciones de un ambiente ideal y (c) las expectativas sobre algún ambiente desconocido.

Cada contexto ambiental se describe en términos de las calificaciones de siete a 10 subescalas que evalúan diversas dimensiones del ambiente, algunas de las cuales se repiten en las escalas que corresponden a contextos diferentes. En efecto, a pesar de la gran diversidad de contextos cubiertos por las Escalas de Clima Social, todas las subescalas de cada contexto se adaptan a la misma clasificación tripartita que incluye (1) dimensiones de relaciones (por ejemplo, compromiso, apoyo, cohesión con los pares), (2) dimensiones del crecimiento personal (digamos, autonomía, orientación a la tarea, competencia) y (3) dimensiones del mantenimiento y cambio del sistema (como, orden y organización, claridad, innovación). También es importante decir que las Escalas de Clima Social se diseñaron para unidades relativamente pequeñas dentro de instituciones complejas y heterogéneas; por ejemplo, un salón de clases más que la escuela entera, un programa de tratamiento en lugar del hospital, un centro de residencia universitaria y no toda la universidad. A este respecto, dichas escalas producen datos menos ambiguos y de más fácil interpretación de los que podrían obtenerse de la evaluación compuesta de organizaciones enteras. Aunque la *Guía del Usuario* de las Escalas del Clima Social (Moos, 1994a) indica que pueden aplicarse en la evaluación clínica, en la consultoría organizacional y en la evaluación de programas, algunos críticos consideran que resultan más adecuadas para la investigación de los determinan-

tes y las consecuencias de las percepciones del ambiente (Allison, 1995; Loyd, 1995; R. O. Mueller, 1995; Saudargas, 1989; Sheehan, 1995; C. R. Smith, 1989).

INFORMES DE LOS OBSERVADORES

Las pruebas que hemos considerado hasta ahora ilustran la diversidad de aproximaciones que se han seguido en la evaluación de la personalidad; con todo, lo más que puede decirse es que constituyen técnicas experimentales prometedoras que son adecuadas para la investigación o instrumentos útiles cuando un clínico experimentado las interpreta junto con otras fuentes de información sobre el sujeto. Está claro que para la evaluación de la personalidad no bastan las pruebas estandarizadas, y que es necesario contar con otras fuentes de información que sigan o complementen las guías proporcionadas por las puntuaciones de las pruebas para evaluar los rasgos para los que no se dispone de instrumentos adecuados y para obtener datos de criterio que permitan el desarrollo y la validación de los tests de personalidad.

Las observaciones directas de la conducta cumplen una parte esencial en la evaluación de la personalidad, ya sea en la clínica, el centro de consejería, el salón de clases la oficina o cualquier otro contexto en el que se requiera de evaluaciones individuales. Para colocar las observaciones conductuales en la perspectiva apropiada, debemos recordar que todas las pruebas son evaluaciones de pequeñas muestras de conducta que se obtienen y evalúan en condiciones estandarizadas; sin embargo, pese a las evidentes ventajas de esos procedimientos estandarizados, debemos equilibrar las ventajas de un muestreo más extenso de la conducta disponible mediante técnicas de observación en medios naturales. Para poner un ejemplo extremo, si tuviéramos la biografía detallada de una persona desde el nacimiento hasta los 30 años probablemente podríamos predecir su conducta subsecuente con mayor precisión de la que nos permitiría cualquier prueba o batería de pruebas. De seguro sería difícil conseguir semejante registro de todas las minucias y circunstancias de la vida, pero si lo tuviéramos podríamos hacer predicciones a partir de una muestra de conducta de 30 años en lugar de hacerlo según las muestras de una o dos horas que nos proporcionan las pruebas.

En todas las técnicas que consideraremos en esta sección, uno o más observadores transmiten lo que hace el individuo en contextos naturales por periodos relativamente largos. A este respecto, es mucho lo que puede hacerse para mejorar tanto la precisión como la comunicación de las observaciones.

Observación naturalista. Los psicólogos infantiles son quienes más han empleado las técnicas de observación directa del comportamiento espontáneo en ambientes naturales, particularmente con preescolares. Aunque los procedimientos pueden ser utilizados con personas de cualquier edad, entre más joven sea el individuo menos probable es que la conducta de interés se vea afectada por la presencia del observador o que el individuo haya desarrollado apariencias sociales que complican la interpretación de la conducta. Las técnicas de observación han demostrado ser útiles en el salón de clases, sobre todo si el observador es el maestro o una persona que se ajuste con facilidad al ambiente normal de la escuela. Las técnicas de evaluación se aplican de manera importante en los programas de modificación de conducta realizados en escuelas, hogares, centros de cuidado infantil, clínicas, hospitales o cualquier otro contexto

(Hartmann y Wood, 1990; Kent y Foster, 1977; Lalli y Goh, 1993), lo mismo que en el campo de la psicología social (Webb, Campbell, Schwartz, Sechrest y Grove, 1981) y en el de la investigación transcultural (Bochner, 1986).

La observación naturalista comprende una gran variedad de procedimientos (Adler y Adler, 1994; Jones, Reid y Patterson, 1975; Sattler, 1988, capítulo 17), que van de las técnicas comprensivas a largo plazo, como el método del diario, a observaciones más restringidas, breves y controladas, como el muestreo cronológico. Este último incluye una distribución representativa de periodos breves de observación cuya duración puede variar entre menos de un minuto y varias horas, dependiendo de la naturaleza y el propósito de la observación, aunque son más comunes los lapsos de cinco minutos o menos. Las observaciones pueden realizarse en un solo día o espaciarse a lo largo de varios meses, y abarcan toda la conducta que ocurra en el periodo especificado; pero a menudo se limitan a una clase de comportamiento, como el lenguaje, la locomoción, la conducta interpersonal o la agresión. Las listas de verificación que indican lo que se debe observar resultan una ayuda valiosa. Otros procedimientos de apoyo incluyen programas de observación, formas de registro, sistemas de codificación y aparatos mecánicos de registro (W. W. Tryon, 1985, capítulos 7 y 8). En ocasiones es posible grabar o filmar los registros[19] y también se dispone de sistemas de microcomputadoras que permiten mejorar la obtención y análisis de los datos de observación (Kratochwill, Doll y Dickson, 1991, pp. 137–141; Repp y Felce, 1990).

Debe advertirse que las observaciones naturalistas tienen mucho en común con las pruebas situacionales que hemos visto, aunque difieren en particular en el hecho de que en las observaciones naturalistas no se ejerce control sobre la situación de estímulo y en que en la mayor parte de los métodos observacionales se registra una muestra más extensa de conducta. El interés en la investigación sobre las técnicas de observación naturalista está en auge, especialmente en relación con su validez ecológica y su utilidad en la evaluación de los cambios a lo largo del tiempo (véase, por ejemplo, Barkley, 1991; Kaminer, Feinstein y Seifer, 1995).

La entrevista. También debemos mencionar la venerable fuente de información de las técnicas de la entrevista que se utilizan con distintos propósitos en la psicología clínica, la consejería, la psicología personal y la educación. En muchas partes se puede encontrar revisiones de los métodos, las aplicaciones y la eficacia de la entrevista, así como de la investigación sobre el proceso de la misma.[20] En cuanto a la forma, las

[19] Un ejemplo inusualmente comprensivo de un sistema de evaluación basado en particular en los datos observacionales es el que formuló Gottman (1994, 1996) para el análisis de los procesos de relaciones matrimoniales. El sistema, que incluye elementos reminiscentes de la metodología del centro de evaluación, se desarrolló y probó durante un programa a largo plazo de investigación destinado a aislar los factores que predicen la posible consecuencia del matrimonio.

[20] La extensa bibliografía sobre intervenciones incluye revisiones periódicas de la investigación correspondiente (Eder, Kacmar y Ferris, 1989; Graves, 1993; Groth-Marnat, 1990, capítulo 3; Landy *et al.*, 1994; McDaniel, Whetzel, Schmidt y Maurer, 1994), así como guías para conducir las entrevistas y recomendaciones para mejorar sus procedimientos, especialmente en la práctica clínica (Bierman, 1990; Lukas, 1993; Morrison, 1995; Rogers, 1995; Shea, 1988) y en la selección de personal (Fear y Chiron, 1990; Webster, 1982). El lector puede encontrar en Gorden (1992) un modelo para la enseñanza y el aprendizaje de las habilidades básicas que pueden aplicarse en la mayor parte de los medios de entrevistas.

variantes van de la entrevista muy estructurada (que representa poco más que un cuestionario aplicado oralmente) y la guiada que cubre áreas predeterminadas, a la entrevista no dirigida y de profundidad, en la que el entrevistador se concreta a establecer la escena y animar al entrevistado para que hable con tanta libertad como le sea posible. Actualmente es común el uso de entrevistas estructuradas para propósitos clínicos y de investigación en el área del diagnóstico psiquiátrico; se trata de instrumentos estandarizados que además de las clasificaciones diagnósticas suelen proporcionar puntuaciones cuantitativas, por lo que tienen que ser evaluadas con los mismos estándares psicométricos de confiabilidad y validez que se aplican a todas las pruebas. El lector interesado en la revisión de protocolos de entrevistas estructuradas puede consultar los trabajos de Hodges y Zeman (1993), Kamphaus y Frick (1996, capítulo 12), y Rogers (1995).

Las entrevistas proporcionan principalmente dos clases de información. Primero, ofrecen una oportunidad óptima para hacer una observación directa de una muestra más bien limitada de la conducta presentada durante la sesión; por ejemplo, puede tomarse nota del habla del individuo, su uso del lenguaje, su aplomo y sus modales al reunirse con un desconocido. Sin embargo, otra función, quizá más importante, consiste en la obtención de datos de la historia del individuo, ya que lo que éste ha hecho en el pasado es un buen indicador de lo que puede hacer en el futuro, en especial cuando se interpreta teniendo en cuenta las circunstancias concomitantes y sus propios comentarios sobre sus acciones. La entrevista no sólo debe interesarse en lo que ha sucedido al individuo, sino también en la forma en que éste percibe y evalúa los acontecimientos.

La entrevista exige de parte del entrevistador la habilidad para obtener e interpretar los datos. Si no obtiene información importante o si no interpreta de modo adecuado la información, la entrevista puede llevarlo a tomar decisiones erróneas. Una cualidad crucial del entrevistador exitoso es su sensibilidad para identificar las señales en la conducta del entrevistado o en los hechos que informa, señales que pueden conducir a un examen posterior de otros hechos que apoyen o contradigan las hipótesis originales.

Estimaciones. Aunque es posible obtener estimaciones en muchos contextos y para distintos propósitos, en esta sección nos interesamos únicamente en el uso de las estimaciones que como evaluación del individuo hace el evaluador a partir de las observaciones acumulativas y no controladas de la vida cotidiana. Estas estimaciones difieren de la observación naturalista en el hecho de que los datos se acumulan de manera casual e informal y en que incluyen la interpretación y juicio en lugar del simple registro de las observaciones; sin embargo, a diferencia de las entrevistas y las observaciones naturalistas, suelen cubrir un periodo de observación más largo y obtienen la información en condiciones más realistas. Las estimaciones se utilizan extensamente en la evaluación de los individuos en ambientes educativos e industriales, en la obtención de datos de criterio para la validación de pruebas y para muchos otros propósitos de investigación. Desde los años setenta ha resurgido la investigación sobre las técnicas de estimación, con énfasis en las investigaciones amplias y sistemáticas y con la suficiente estandarización de definiciones y procedimientos que faciliten la comparación de los descubrimientos de diversos estudios (Borman, 1991; Landy y Farr, 1983; Ozer y Reise, 1994, pp. 370–371; Saal, Downey y Lahey, 1980; Sulsky y Balzer, 1988).

Es mucho lo que puede hacerse para mejorar la exactitud de las estimaciones. Un problema común surge de la ambigüedad en el nombre de los rasgos, las unidades de la escala o ambos. Para enfrentar esta dificultad, cada rasgo debe definirse en términos específicos, y las estimaciones tienen que expresarse en tal forma que todos los evaluadores las interpreten de manera uniforme. En lugar de usar números o adjetivos generales que transmiten diferentes significados a calificadores distintos, es necesario identificar con mayor claridad los grados de cada rasgo en términos de soportes conductuales cuidadosamente formuladas (Dickinson y Zellinger, 1980). También hay evidencias de que la relativa precisión de los formatos de diferentes escalas puede variar con la naturaleza del trabajo o la función que se quiere calificar (Borman, 1979; J. M. Feldman, 1986).

Una de las condiciones que afectan la validez de las estimaciones es el grado de *contacto relevante* que el calificador tiene con la persona que debe valorarse (Freeberg, 1969; Landy y Farr, 1980; Paulhus y Bruce, 1992; Wiggins y Pincus, 1992, pp. 493–496). No basta con conocer al individuo por mucho tiempo, el evaluador debe haber tenido la oportunidad de observarlo en situaciones en que pudiera manifestarse la conducta en cuestión; por ejemplo, si un empleado nunca ha tenido la oportunidad de tomar decisiones en su trabajo, el supervisor no puede evaluar su capacidad para hacerlo. En muchas situaciones, es preferible incluir un espacio para ser supervisado en lugar de una estimación si el calificador no ha tenido la oportunidad de observar cierto rasgo en el individuo.

Las estimaciones son susceptibles de muchos errores.[21] Un ejemplo bien conocido es el *efecto de halo*, un fenómeno que tradicionalmente se ha definido como la tendencia de los calificadores a dejarse influir por un solo rasgo favorable o desfavorable que matiza el juicio que hagan de los otros rasgos. Por lo común, la presencia del efecto de halo se infiere de las correlaciones observadas entre las estimaciones asignadas en diferentes dimensiones de la ejecución. Aunque algunos investigadores lo consideran todavía de la misma manera, y como un constructo unitario, han destacado otras formas posibles de concebirlo y de definirlo operacionalmente (Balzer y Sulsky, 1992; Kozlowski, Kirsch y Chao, 1986; Murphy y Anhalt, 1992; Nathan, 1986).[22]

Otros errores que pueden afectar los datos de la estimación son el *error de la tendencia central* (la tendencia a colocar a las personas en el medio de la escala y evitar las posiciones extremas) y el *error de indulgencia*, que se refiere a la renuencia de muchos evaluadores a asignar estimaciones desfavorables. El primero ocasiona que las calificaciones se agrupen en el centro de la escala; el segundo, que se agrupen en el extremo superior. Estos errores pueden eliminarse mediante el uso de un ordenamiento u otros *procedimientos de orden por mérito* que obligan a establecer una discriminación entre individuos y, por ende, a maximizar la información generada por las estimaciones. Por supuesto, las técnicas que sean aplicables cuando se hacen comparaciones dentro del mismo grupo no permiten comparaciones directas entre grupos evaluados por calificadores diferentes.

Un hallazgo inesperado de la investigación acumulada de los errores de estimación es que la relación entre las medidas de errores, como los de halo o de indulgencia y otros indicadores más directos de la precisión de las estimaciones, está lejos de ser sen-

[21] Gran parte de la investigación sobre los errores en la valoración se informa en el *Journal of Applied Psychology*.

[22] Algunas de estas otras conceptuaciones consideran al error de halo, al menos en parte, como una función del calificado, e incluso en el contexto específico de la evaluación, así como del calificador.

cilla, y a menudo va en contra de la intuición. Varios investigadores que han realizado estudios meta-analíticos y que han revisado la bibliografía sobre los errores de estimación han llegado a la conclusión de que los problemas metodológicos y conceptuales inherentes a la apreciación del desempeño mediante estimaciones impide el uso de las medidas de estos errores como indicador de su precisión (Balzer y Sulsky, 1992; Borman, 1991; Murphy y Anhalt, 1992; Murphy y Balzer, 1989).

No obstante, el proceso de estimación suele mejorar con la capacitación de los calificadores. En diversos medios se ha demostrado la eficacia de dichos procedimientos para aumentar la confiabilidad y la validez de las estimaciones y para reducir los errores comunes de juicio (Bernardin y Buckley, 1981; McIntyre, Smith y Hassett, 1984; Pulakos, 1986; Stamoulis y Hauenstein, 1993; Sulsky y Day, 1992, 1994); sin embargo, debe advertirse que en los programas de capacitación de evaluadores se han incluido muchos tipos diferentes de entrenamiento cuyos efectos varían en clase, cantidad y duración. La capacitación puede consistir en dar a los calificadores una norma estándar de referencia que les permita evaluar el desempeño, analizar los errores comunes en la estimación y las formas de minimizar su influencia; aunque otras formas de capacitación pueden ser más apropiadas para contextos y propósitos particulares, es probable que en la mayor parte de las situaciones la mejora de las habilidades de observación de los calificadores arroje resultados favorables.

La evaluación clínica a menudo requiere la obtención de datos de informantes que están familiarizados con los patrones conductuales del individuo que está siendo evaluado. Las escalas de estimación proporcionan una forma eficaz de conseguir esos datos y son útiles para obtener estimaciones de niños y adolescentes. En los últimos años, se ha publicado y está comercialmente disponible en los Estados Unidos, un gran número de escalas estandarizadas para obtener estimaciones de padres y maestros.[23] Un ejemplo particularmente amplio es el Sistema de Evaluación Conductual para Niños (*Behavior Assessment System for Children*, BASC; Reynolds y Kamphaus, 1992). El sistema incluye componentes autodescriptivos y datos de un observador además de las escalas de estimación del maestro y las de los padres, de las que existen tres versiones que van de preescolar a la adolescencia (para una revisión, véase R. B. Kline, 1994).

Técnica nominativa. Un procedimiento muy útil para obtener estimaciones de los compañeros es la técnica nominativa. Aunque originalmente fue desarrollada en la sociometría (J. L. Moreno, 1953) para investigar la estructura de los grupos, puede utilizarse en cualquier grupo de personas que hayan estado juntas durante el tiempo suficiente para conocerse, como los salones de clases, las fábricas, los clubes o las unidades militares. Se pide a cada individuo que elija a uno o más miembros del grupo con los que le gustaría estudiar, trabajar, hacer el almuerzo, jugar o desempeñar cualquier otra tarea; se le puede pedir que mencione a tantos miembros del grupo como desee, un número específico (por ejemplo, la primera, segunda o tercera elección) o sólo a una persona para cada función.

Cuando se utilizan para hacer evaluaciones individuales, las nominaciones recibidas por una persona pueden servir para identificar a líderes potenciales (quienes reciben

[23] El lector puede encontrar en Kamphaus y Frick (1996), Piacentini (1993) y Witt, Heffer y Pfeiffer (1990) revisiones de este tipo de instrumento y de algunos ejemplos concretos.

muchas elecciones) y a los individuos aislados (los que rara vez son mencionados). Además, para hacer una evaluación más precisa de cada individuo pueden calcularse varios índices, de los que el más simple es el conteo del número de veces que el individuo fue nominado para una función específica, lo que puede considerarse como una estimación de sus compañeros. La técnica nominativa puede emplearse en referencia a cualquier conducta de interés; por ejemplo, puede pedirse a los examinados que nombren a la persona que tiene las ideas más originales, en quién puede confiarse que cumplirá con el trabajo o quién es el mejor deportista. Asimismo, se les puede pedir que designen a la persona que más se parece y a la menos parecida a la descripción proporcionada; en este caso, las nominaciones positivas recibirán un peso de +1 y las nominaciones negativas se contarán como –1 para obtener la puntuación total de cada persona.[24] Debe agregarse que las evaluaciones de los pares también pueden obtenerse mediante otros procedimientos, como el ordenamiento o la estimación, pero la técnica nominativa es la que ha demostrado ser más exitosa y, por consecuencia, la que se emplea con mayor frecuencia.

Como quiera que se obtengan, las estimaciones de los compañeros generalmente resultan ser una de las técnicas de valoración más confiables en grupos tan diversos como el personal militar, los supervisores industriales, los voluntarios de los cuerpos de paz, niños escolares y universitarios (Cole y White, 1993; Gresham y Little, 1993; Hughes, 1990; Kamphaus y Frick, 1996, capítulo 10; Kane y Lawler, 1978; J. S. Wiggins, 1973/1988, pp. 356–363). Cuando se comparan con diversos criterios prácticos que dependen de las relaciones interpersonales, es común encontrar que la validez concurrente y predictiva de las estimaciones es buena, lo que resulta comprensible cuando consideramos algunas de las estimaciones de los rasgos hechas por los compañeros. Primero, el número de calificadores suele ser grande, pues incluye a todos los miembros del grupo. Segundo, los compañeros del individuo a menudo se encuentran en una posición favorable para observar su conducta, por lo que pueden resultar mejores jueces de ciertos rasgos que maestros, supervisores y otros observadores externos. Tercero, y quizá más importante, las opiniones de los miembros del grupo —correctas o erróneas— influyen en las acciones del individuo y por ende determinan parcialmente la naturaleza de sus interacciones subsecuentes con el grupo. Es de esperarse que otros grupos comparables reaccionen con el individuo de manera similar, de ahí que pueda afirmarse que las estimaciones sociométricas tienen validez de contenido en el mismo sentido que las muestras de trabajo.

Listas de verificación y clasificaciones Q. Al describir a otra persona, el observador puede emplear cualquier instrumento autodescriptivo, como los inventarios de intereses y de personalidad que explicamos en los capítulos 13 y 14.[25] Las pruebas diseñadas para evaluar el autoconcepto resultan especialmente convenientes para ese

[24] Cuando se obtienen nominaciones negativas, los examinadores deben tener especial cuidado de impedir cualquier efecto potencialmente nocivo del procedimiento. Para una revisión de la ética de las evaluaciones referidas a los pares con niños y adolescentes, véase Gresham y Little (1993, pp. 174–175) y Kamphaus y Frick (1996, pp. 201–203).

[25] El Inventario de Personalidad para Niños (*Personality Inventory for Children*, PIC) y el Inventario de Personalidad para Jóvenes (*Personality Inventory for* Youth, PIY) —que estudiamos en el capítulo 13— son, en efecto, instrumentos paralelos en (a) la forma autodescriptiva (PIY), y (b) en la forma de un inventario de conducta observada (PIC).

propósito. En el programa de investigación del IPAR es común el uso de la Lista de Verificación de Adjetivos (*Adjective Check List*, ACL) para obtener evaluaciones de los observadores (Gough y Heilbrun, 1983); en este caso, los psicólogos que han observado de cerca al participante por un periodo de evaluación de dos o tres días registran sus evaluaciones verificando en la lista los adjetivos adecuados.

Otro instrumento común en las evaluaciones del observador es la clasificación Q. Block (1961/1978) desarrolló originalmente el Inventario de Clasificación Q de California (*California Q-Sort Deck*) para proporcionar a los observadores profesionales un lenguaje común para sus evaluaciones de la personalidad. Después se publicó el inventario para una distribución más general y se revisó para que su lenguaje fuera apropiado para legos y profesionales. Asimismo, se puso en circulación una adaptación para los más jóvenes, el Conjunto Q de California para Niños (*California Child Q-Set*; para una revisión, véase Heilbrun, 1985). En todas sus formas, el material consta de 100 afirmaciones que deben clasificarse en una distribución forzada de nueve puntos y que se categorizan de acuerdo con su "relevancia" para el individuo, es decir, su importancia para especificar sus características únicas y esenciales, lo que supone que se conserva el marco ipsativo de referencia que es característico de las clasificaciones Q; así, el individuo no se compara con estándares normativos externos.

Contar con conjuntos uniformes de reactivos de clasificación Q facilita la comunicación y asegura la comparabilidad de los datos de distintos observadores. La clasificación Q también puede utilizarse para otros propósitos de investigación (véase, por ejemplo, Caspi *et al.*, 1992; Reise y Oliver, 1994; Wink, 1992). Otra aplicación incluye su uso en las evaluaciones individuales. A ese respecto, Block (1961/1978) proporciona ejemplos de tres "definiciones de clasificaciones Q" que representan una evaluación por consenso de un individuo completamente normal y de dos síndromes psiquiátricos, con los que pueden compararse clasificaciones Q individuales. Además, es posible desarrollar clasificaciones similares para cualquier categoría de personas.

DATOS BIOGRÁFICOS

En la revisión que hicimos de las técnicas de la entrevista hablamos de la importancia de contar con los datos de la historia del sujeto. La información acerca de la conducta y las experiencias del individuo resulta de interés tanto para los teóricos que tratan de comprender los patrones de personalidad y el desarrollo cognoscitivo como para los expertos en psicología aplicada que pretenden evaluar y predecir la conducta de los individuos. Esto obedece principalmente a que la forma en que el individuo ha respondido en situaciones concretas es una fuente prometedora de información respecto al modo en que responderá a situaciones similares en el futuro.

Los datos de la historia vital pueden obtenerse por diversos métodos, de los cuales la entrevista y los cuestionarios son tal vez los más comunes, sobre todo en la consejería y en la psicología clínica. Los datos y escritos autobiográficos también proporcionan una fuente rica de información para los psicobiógrafos y otros estudiosos interesados en el estudio de la vida de los individuos (véase, por ejemplo, J. S. Wiggins y Pincus, 1992, pp. 487–493).[26] Por su parte, los investigadores que realizan estu-

[26] Estudiamos el uso de los recuerdos autobiográficos como instrumento proyectivo en el capítulo 15.

dios longitudinales no sólo reúnen, sino también crean registros de los datos biográficos mediante las observaciones y mediciones repetidas de sus sujetos a lo largo del tiempo (Funder, Parke, Tomlinson-Keasey y Widaman, 1993).

Sin embargo, el método más estructurado para obtener y utilizar estos datos son las escalas o inventarios biográficos (conocidos colectivamente como medidas de los datos biográficos o biodata), que se formulan para predecir el desempeño en ambientes industriales y educativos. Al igual que los inventarios de intereses y de personalidad que vimos en los capítulos 13 y 14, el inventario biográfico es un instrumento estandarizado autodescriptivo en el que el examinado no redacta sus respuestas a los reactivos, sino que las selecciona de entre dos o más opciones. Aunque casi todos los cuestionarios atañen a hechos verificables y relativamente objetivos, cuando se califican y emplean para la evaluación o predicción, las escalas biográficas deben cumplir, al igual que las entrevistas estructuradas, con los mismos criterios psicométricos de confiabilidad y validez que cualquier otro instrumento. Los reactivos característicos suelen abordar la cantidad y la naturaleza de la educación, la experiencia laboral, las habilidades especiales, los pasatiempos y las actividades recreativas. Con frecuencia se busca la reacción del individuo a sus experiencias, como cuando se le pregunta acerca de los cursos que más y menos le agradaban en la escuela o lo que le gustaba o no de sus experiencias laborales.[27]

Históricamente, los reactivos de la escala biográfica han sido seleccionados y sopesados mediante una clave empírica de criterio, como en la elaboración de inventarios como el MMPI y el Strong, que revisamos en los capítulos 13 y 14. El inventario resultante se somete luego a una validación cruzada contra el mismo criterio en una nueva muestra. Cuando se siguen esos procedimientos, los inventarios de datos biográficos han demostrado de manera constante que son buenos predictores del desempeño en diversos contextos. Se han desarrollado inventarios comparados con criterios tan variados como la cantidad de pólizas vendidas por agentes de seguros de vida, los cambios de trabajo de empleados bancarios, la productividad de investigadores científicos, la creatividad artística de estudiantes de secundaria y el desempeño del personal naval en diversos entrenamientos. Estas medidas han demostrado ser predictores válidos del desempeño en el trabajo de grupos que van de los obreros no calificados a los profesionales y ejecutivos de alto nivel (Anastasi, 1979, pp. 79–80; Owens, 1983). Al mismo tiempo, muchos inventarios biográficos preparados para trabajos concretos mediante métodos empíricos tienen la limitación de no ser transportables o generalizables, por lo que su aplicación tiende a limitarse a los ambientes y los criterios particulares a los que fueron destinados (Baird, 1985; Hunter y Hunter, 1984).

Las medidas biográficas aún son objeto de una extensa investigación. Se ha realizado una minuciosa exploración mediante análisis factoriales y racionales del desarrollo de las escalas de datos biográficos (Hough y Paullin, 1994; Schoenfeldt y Mendoza, 1994). En constraste con la selección de la clase de criterio empírico de los reactivos, la aproximación racional suele comenzar con la identificación de los constructos más relevantes por medio del análisis de la obra y la revisión de las fuentes publicadas en la bibliografía teórica y empírica. A esto sigue el análisis factorial del

[27] El lector puede encontrar en Mael (1991) una presentación minuciosa del dominio y los atributos de los datos biográficos o biodata.

banco de reactivos preliminar, del que se seleccionan los reactivos finales para las escalas del factor resultante. Los méritos relativos de las diversas aproximaciones en la elaboración de las escalas de datos biográficos están sujetos a un continuo debate, aunque está claro que cada una tiene sus puntos fuertes así como sus debilidades. La aproximación ideal es la estrategia de validación de constructo que incorpore tantos elementos diversos como sea posible (Hough y Paullin, 1994).

Además de las aproximaciones tradicionales al desarrollo del inventario, se están ensayando nuevos métodos para generar, seleccionar y establecer la clave de los reactivos de los datos biográficos en un intento por aumentar la generabilidad de los instrumentos resultantes (véase, por ejemplo, Rothstein, Schmidt, Erwin, Owens y Sparks, 1990; Russell, Mattson, Devlin y Atwater, 1990). William A. Owens fue pionero de una de las líneas de investigación más productivas que utilizan las técnicas de agrupamiento para identificar los subgrupos de individuos que comparten patrones modales de experiencias personales (Mumford, Stokes y Owens, 1990; Mumford y Stokes, 1992; Owens y Schoenfeldt, 1979). Esos descubrimientos pueden aplicarse luego a la predicción de criterios múltiples. Más importante, el estudio de los subgrupos identificados con estos métodos puede llevar a un nivel de comprensión de los patrones de desarrollo en la vida de los individuos que combine los elementos de las aproximaciones idiográfica y nomotética (Hein y Wesley, 1994).

En la actualidad, luego de años de una investigación cuidadosamente controlada, las mediciones de los datos biográficos se encuentran entre los promedios más confiables y eficaces para la evaluación y la selección en la educación, la industria, el gobierno y otros contextos. Más aún, la historia de la investigación sobre los datos biográficos resalta de manera clara la mutua dependencia de las ciencias básicas y las aplicadas. Es decir, los métodos y resultados de los estudios realizados para resolver problemas pragmáticos de selección de personal están contribuyendo a la formulación de un marco teórico de referencia para comprender el desarrollo de patrones de conducta a lo largo del ciclo vital.

No obstante, poner en práctica las pruebas de datos biográficos no está exento de problemas técnicos, prácticos y políticos; entre ellos destacan la preocupación por (a) temas legales como la invasión de la privacidad y asuntos de igualdad de oportunidades de empleo (Sharf, 1994), y (b) la susceptibilidad de las mediciones de los datos biográficos y los informes retrospectivos del falseamiento y otras fuentes de inexactitud (Henry, Moffitt, Caspi, Langley y Silva, 1994; Lautenschlager, 1994; Trent y Laurence, 1993). Buena parte de la investigación disponible, sobre todo los aspectos del desarrollo y uso de los instrumentos biográficos, ha sido resumida en el *Biodata Handbook* (Stokes, Mumford y Owens, 1994).

QUINTA PARTE

Aplicaciones de las pruebas

Principales contextos del uso actual de las pruebas

Las pruebas psicológicas se utilizan con diversos propósitos y sus áreas de aplicación crecen continuamente. Luego de haber revisado con cierto detalle ejemplos representativos de distintas pruebas, pasemos a la consideración de los temas que atañen a su aplicación. En este capítulo consideramos tres áreas principales en las que las pruebas cumplen una multiplicidad de funciones, que a grandes rasgos podemos llamar los contextos educativo, ocupacional y clínico o de consejería. En el siguiente capítulo, el último, veremos los temas éticos y sociales que se relacionan con las prácticas de evaluación en todas las áreas.

EVALUACIÓN EDUCATIVA

En las escuelas se emplea casi cualquier prueba disponible; en el repertorio del consejero y de los psicólogos escolares pueden encontrarse tests de inteligencia, de aptitudes especiales, de aptitudes múltiples y de personalidad. Es común que maestros y directores tengan que actuar en función de los resultados obtenidos con diferentes clases de pruebas; sin embargo, se han desarrollado específicamente algunos instrumentos para emplearse en los contextos educativos,[1] pruebas que constituyen el centro de interés de esta sección. Incluyen instrumentos diseñados para la predicción y clasificación en ambientes educativos y muchas clases de pruebas de aprovechamiento escolar.

[1] Tanto la versión de 1985 de los *Estándares de Examinación* (AERA, APA, NCME) como su revisión propuesta (véase el capítulo 1) incluyen un capítulo dedicado al uso de las pruebas en la educación.

Naturaleza y usos de las pruebas de aprovechamiento. Las pruebas de *aprovechamiento*, cuyo número supera con mucho a cualquier otro tipo de pruebas estandarizadas, están diseñadas para medir los efectos de los programas de instrucción o entrenamiento. Se acostumbra distinguirlas de los tests de aptitud, que incluyen varios tests generales de inteligencia, baterías de aptitudes múltiples y tests de aptitudes especiales. Desde un punto de vista, la diferencia entre la evaluación del aprovechamiento y la de la aptitud estriba en el grado de uniformidad de las experiencias relevantes. De este modo, una prueba de aprovechamiento mide los efectos de conjuntos relativamente estandarizados de experiencias, como un curso elemental de francés, de trigonometría o de programación en computadoras. En contraste, la ejecución en un test de *aptitud* refleja la influencia acumulativa de una multiplicidad de experiencias en la vida diaria. Podemos decir que los tests de aptitud miden el efecto del aprendizaje en condiciones relativamente no controladas y desconocidas, mientras que las pruebas de aprovechamiento miden los efectos del aprendizaje ocurrido en condiciones que en parte se conocen y controlan.

La segunda distinción entre los tests de aptitud y las pruebas de aprovechamiento tiene que ver con sus usos; los primeros permiten predecir la ejecución subsecuente, se emplean para estimar el grado en que los individuos se beneficiarán de un curso específico de entrenamiento o para anticipar la calidad de su aprovechamiento en una nueva situación. Por otro lado, las pruebas de aprovechamiento generalmente representan una evaluación terminal de su posición individual al finalizar la formación. En éstas, el énfasis se centra en lo que el individuo puede hacer en ese momento.

Sin embargo, debe reconocerse que no es posible aplicar rígidamente distinción alguna entre los tests de aptitud y las pruebas de aprovechamiento. Algunos tests de aptitud pueden depender de un aprendizaje muy específico y uniforme, mientras que algunas pruebas de aprovechamiento cubren experiencias educativas más bien amplias y no estandarizadas. De modo similar, una prueba de aprovechamiento puede emplearse para predecir el aprendizaje futuro y, como tal, cumple la misma función que un test de aptitud; por ejemplo, las pruebas de aprovechamiento en los cursos propedéuticos de la escuela de medicina pueden servir para predecir el rendimiento en esos estudios.

En un esfuerzo por evitar el significado excesivo que se ha asociado a los términos "aptitud" y "aprovechamiento", cada vez es más frecuente el uso del término "habilidad" para designar las mediciones de la conducta cognoscitiva.[2] Cualquier prueba cognoscitiva, independientemente de cómo se llamara antes, ofrece una muestra de lo que el individuo sabe en el momento en que se le probó y mide el nivel de desarrollo obtenido en una o más habilidades. Ningún instrumento revela cómo o por qué alcanzó ese nivel, y para responder a esas preguntas es necesario ahondar en otras variables concomitantes y en especial en los antecedentes y las experiencias del individuo. En

[2] Los nuevos nombres de los exámenes de la Junta Universitaria, efectivos desde 1994, representan un ejemplo de los cambios que tienen lugar en relación con las etiquetas. La conocida abreviatura del SAT ahora corresponde a Test de Evaluación Escolar (*Scholastic Assessment Test*) en lugar de Test de Aptitud Académica (*Scholastic Aptitude Test*). El nuevo SAT se ha reagrupado en dos componentes, el SAT-I, una prueba de razonamiento que reemplaza al anterior Test de Aptitud Académica, y el SAT-II, son pruebas temáticas que reemplazan a las anteriores pruebas de aprovechamiento. Estos cambios fueron acompañados por otras innovaciones sustanciales en las pruebas que veremos en otra sección del capítulo.

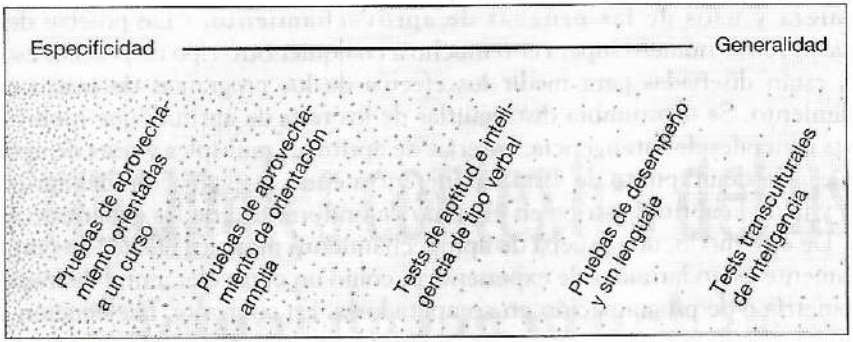


Figura 17.1. Pruebas de habilidades desarrolladas: continuo de especificidad de la experiencia.

este sentido, cada puntuación tiene un pasado que debe ser explorado para la adecuada comprensión del sujeto que la obtuvo; pero también tiene un futuro en la medida en que permite cierta predicción de lo que hará en otras situaciones, ajenas a la prueba, así como en algún tiempo futuro.

Los tests de habilidades desarrollados, se han diseñado como tests de inteligencia general, baterías de aptitudes múltiples, tests de aptitudes especiales o pruebas de aprovechamiento, pueden ordenarse en términos de la especificidad de la experiencia que suponen, como se ilustra en la figura 17.1. En un extremo se encuentran las pruebas de aprovechamiento orientadas al curso que cubren habilidades técnicas o información real muy definidas. Aquí podría encontrarse, por ejemplo, un examen de vocabulario de ruso o de mantenimiento de televisores. Vienen luego las pruebas de aprovechamiento de orientación más amplia, que suelen usarse para evaluar la obtención de las principales metas educativas a largo plazo, y en las que encontramos instrumentos que se concentran en la comprensión y aplicación de principios científicos, la interpretación de la literatura o la apreciación de las artes. De orientación aún más amplia son los instrumentos de medición de las habilidades cognoscitivas básicas, como las pruebas de comprensión de lectura, de cálculos aritméticos y de razonamiento lógico, que afectan el rendimiento del individuo en una gran variedad de actividades. Es claro que aquí se funden imperceptiblemente las pruebas de aprovechamiento con los instrumentos de medición tradicionales de inteligencia y de aptitud.[3] Las baterías cognoscitivas predominantemente verbales, conocidas como tests de inteligencia, están próximas a las pruebas más amplias de aprovechamiento; vienen luego las de ejecución y sin lenguaje, que por lo general no requieren de lectura o escritura; mientras que en el extremo se encuentran los instrumentos transculturales diseñados para emplearse con personas de muy diversos antecedentes.

Denominar a algunos instrumentos como "tests de aptitud" y a otros "pruebas de aprovechamiento" puede dar lugar a algunos usos inadecuados de los resultados. Un

[3] Esta superposición puede demostrarse empíricamente mediante el examen del cercano parecido del contenido y el grado de correlación entre las pruebas con esas etiquetas (véase, por ejemplo, W. Coleman y Cureton, 1954; Cooley y Lohnes, 1976).

ejemplo común es la práctica de identificar como niños con bajo aprovechamiento a aquellos cuyas puntuaciones en las pruebas de aprovechamiento son inferiores a los obtenidos en los tests de aptitud o en los de inteligencia. En realidad, estas diferencias de los individuos en las puntuaciones de las pruebas reflejan el hecho de que no hay dos instrumentos (u otros indicadores de la ejecución, como las calificaciones obtenidas en los cursos) que mantengan una correlación perfecta. La cuestión del bajo o elevado aprovechamiento puede formularse con mayor precisión como una sobre o subpredicción de la primera prueba a la segunda. Entre las razones que explican los errores de predicción en los casos individuales se encuentran la poca confiabilidad de los instrumentos, las diferencias en el contenido cubierto, los diversos efectos de los factores actitudinales y motivacionales sobre ambas medidas y el impacto de experiencias como la instrucción de remedio o una larga enfermedad (R. L. Thorndike, 1963).

Desde hace mucho tiempo se reconocen las diversas funciones que pueden cumplir las pruebas de aprovechamiento en el proceso educativo. Como auxiliar en la asignación de calificaciones (o en cualquier otra evaluación de la competencia alcanzada), las pruebas estandarizadas de aprovechamiento tienen las ventajas de la objetividad, la uniformidad y la eficacia. Si se construyen de manera apropiada, tienen otros méritos, como la adecuada cobertura del contenido y la reducción de la operación de factores aleatorios e irrelevantes en los procedimientos de calificación. Las pruebas de aprovechamiento también constituyen una característica importante de los programas de enseñanza de remedio. A este respecto, pueden ser de utilidad tanto en la identificación de los estudiantes con problemas educativos especiales como en la medición del progreso en el curso del trabajo de remedio. Para todos los estudiantes, la aplicación periódica de pruebas de aprovechamiento bien elaboradas y apropiadamente elegidas pueden ayudar a facilitar el aprendizaje. Éstas pueden revelar las debilidades del aprendizaje anterior, dar dirección al aprendizaje subsecuente y motivar al estudiante. El valor de incentivo del "conocimiento de los resultados" ha sido demostrado repetidamente en algunos experimentos psicológicos realizados en muchas situaciones de aprendizaje, con sujetos que varían mucho tanto en edad como educación. La eficacia de dicha autosupervisión suele aumentar por su inmediatez.

Desde otro punto de vista, las pruebas de aprovechamiento proporcionan un medio para adaptar la instrucción a las necesidades individuales. La enseñanza puede ser más fructífera cuando es adecuada para la etapa del alumno. El primer paso necesario para la enseñanza eficaz consiste en asegurarse de qué es lo que el estudiante puede hacer y qué es lo que ya sabe. Al aplicar exámenes al inicio del año escolar, los educadores pueden emprender los pasos necesarios para cubrir las deficiencias reveladas por los resultados. Otros ejemplos de la función de las pruebas de aprovechamiento en el proceso de enseñanza se aprecian en la evaluación referida a dominio y los sistemas instruccionales individualizados que analizamos en el capítulo 3.

Finalmente, las pruebas de aprovechamiento pueden emplearse como apoyos en la evaluación y el mejoramiento de la enseñanza, así como en la formulación de las metas educativas. Las pruebas de aprovechamiento pueden proporcionar información sobre lo adecuado de la enseñanza de las habilidades y contenido esencial. Al concentrar la atención en estos temas y ofrecer hechos concretos, las pruebas estimulan el análisis de los objetivos educativos y fomentan el examen crítico del contenido y de

los métodos de instrucción.[4] Dado el incremento en la exigencia pública de explicaciones, en las últimas décadas se ha observado una cantidad sin precedente de evaluaciones a instituciones educativas de todos los niveles. En los Estados Unidos, gran parte de esta evaluación es obligatoria, por mandato de los distritos locales y estatales, así como por el gobierno federal (B. Gifford, 1989b; Hartle y Battaglia, 1993; *National Council on Education Standards and Testing*, 1992). La Evaluación Nacional del Progreso Educativo, conocida informalmente como "la tarjeta de calificaciones nacionales", es uno de los ejemplos mejor conocidos del programa de evaluación de ley establecido por el gobierno estadounidense (véase, por ejemplo, Alexander y James, 1987; Gentile, Martin-Rehrmann y Kennedy, 1995; E. G. Johnson, 1992; Messick, Beaton y Lord, 1983; NAEP, 1985; F. B. Womer, 1970). Debido a la difusión de las pruebas estandarizadas y a las consecuencias económicas de amplio rango que pueden estar unidas a sus resultados, los propios instrumentos han sido sometidos a críticas y a un escrutinio intenso. Los métodos utilizados para evaluar el progreso educativo, alguna vez territorio exclusivo de los especialistas en las pruebas, se ha convertido en el centro de interés de un debate muy politizado que no sólo recibe la atención de legisladores y líderes de la industria, sino también del público en general (R. E Bennett y Ward, 1993; Courts y McInerney, 1993; Gifford y O'Connor, 1992; S, P. Robinson, 1993; G. P. Wiggins, 1993). El lector interesado puede encontrar en Linn y Gronlund (1995, capítulo 18) una revisión de los temas y las tendencias relacionados con los exámenes escolares y la evaluación de los programas.

Construcción en comparación con elección. Históricamente, los exámenes escolares tradicionales consistían en un conjunto de preguntas que debían responderse de palabra o por escrito. En cualquier caso, el examinado componía y formulaba la respuesta; el término "pregunta de ensayo" llegó a utilizarse ampliamente para cubrir las preguntas de respuesta libre, que no sólo incluían las que exigían un largo ensayo, sino también las que pedían al examinado dar una respuesta corta o trabajar en la solución de un problema matemático. En contraste, las "preguntas objetivas" eran aquellas que pedían la elección de una respuesta correcta de las alternativas proporcionadas para cada pregunta. Aunque hay varias clases de reactivos que piden a los examinados seleccionar una respuesta, como los de falso o verdadero y los de emparejamiento, con mucho las más utilizadas, estudiadas y también más criticadas han sido las preguntas de opción múltiple.

Los críticos de éstas argumentan que promueven la memorización mecánica y el aprendizaje de hechos aislados en lugar del desarrollo de las habilidades para la resolución de problemas y la comprensión de conceptos. Además, mucha gente mal informada en los planteles y la política educativa considera idénticos los reactivos de opción múltiple con la evaluación estandarizada y menoscaba a la vez estos dos elementos de la metodología de la evaluación.[5] En un giro irónico de los acontecimien-

[4] Publicaciones recientes en la educación en las áreas de matemáticas y ciencia (véase, por ejemplo, Pennner, Batsche, Knoff y Nelson, 1993) y el mejoramiento de las habilidades de pensamiento (véase, por ejemplo, Mulcahy, Short y Andrews, 1991) ejemplifican el trabajo orientado a la solución que puede resultar de este examen crítico.

[5] Debe mencionarse que los instrumentos utilizados en muchos programas de evaluación educativa a gran escala (como las pruebas del SAT, el NAEP y otros instrumentos estandarizados de aprovechamiento) utilizaron por algún tiempo preguntas de ensayo y reactivos de formato abierto.

tos, a menudo se considera que los mismos programas estandarizados de evaluación usados para planear el progreso educativo contribuyen a las deficiencias educativas que han descubierto (Courts y McInerney, 1993; H. Gardner, 1992; Resnick y Resnick, 1992). Por desgracia, las críticas acerca del uso excesivo e inadecuado de las pruebas estandarizadas en algunos casos han sido cuidadosamente justificadas. En cualquier caso, las quejas de que la evaluación dirige el currículum y que ambos necesitan con urgencia una reforma provienen de educadores de todos los niveles. Los defensores de la reforma educativa creen que se necesita una revisión general tanto de las metas curriculares, de los métodos instruccionales así como de los instrumentos de evaluación debido a que perciben que todas esas áreas se encuentran relacionadas

Como los aspectos filosóficos, políticos y prácticos de la reforma educativa están más allá del alcance de este libro, nos limitaremos a analizar algunas de las opciones propuestas en la metodología de la evaluación, las cuales se describen por diversos rubros como evaluación "basada en el desempeño", evaluación "auténtica" y evaluación "directa" (véase, por ejemplo, E. L. Baker, O'Neil y Linn, 1993; Linn y Gronlund, 1995, capítulo diez). Aunque cada una de ellas representa énfasis distintos, una de las características centrales que todos comparten es la marcada preferencia por tareas que, como las anteriores "preguntas de ensayo", requieren que el examinado dé una respuesta. Estos reactivos se describen ahora como de *construcción de respuesta* o *tareas abiertas*, y se distinguen de las *tareas de selección de respuesta*, que es el término aplicado a los reactivos que sólo exigen que el examinado elija una respuesta de entre las opciones presentadas, como los reactivos de opción múltiple y los otros tipos de preguntas que solían llamarse "objetivas". Los reactivos de construcción de respuesta pueden incluir el simple completamiento (como los de "llenar los espacios"), solución de problemas y ensayos, así como una variedad de presentaciones o desempeños como tocar un instrumento, dar una charla o reparar una máquina.[6]

El método conocido como *evaluación de portafolio* ofrece otro conjunto de alternativas. Esta herramienta pretende principalmente hacer tan significativo y realista como sea posible el proceso de evaluación educativa. Aunque el término se aplica a una amplia gama de procedimientos, por lo común el portafolio consta de un registro acumulativo (obtenido a lo largo de un periodo prolongado) de muestras del trabajo del estudiante en áreas concretas, como la redacción o cualquier otro empeño que incluya un proceso en el que sea posible documentar el progreso (Camp, 1993; Gitomer, 1993; D. P. Wolf, 1993). El método de portafolio ofrece gran flexibilidad y puede ponerse en práctica de manera más o menos formal y con diversos grados de colaboración entre el estudiante y el maestro (véase en Karlsen, 1992, un ejemplo de un instrumento publicado).

El lector habrá inferido, incluso de esta breve recapitulación, que se está prestando gran atención a los medios por los que se conducen las evaluaciones del aprendizaje y el trabajo del estudiante. Esta preocupación se extiende no sólo a lo que miden los di-

[6] Obviamente, los reactivos de la prueba pueden clasificarse según otras dimensiones, además de si incluyen o no la construcción o la selección de respuestas. En R. E. Bennett (1993) y en R. E. Snow (1993) pueden encontrarse ejemplos de dos diferentes taxonomías de los tipos de reactivos.

ferentes reactivos y a lo bien que lo hacen, sino también a otros aspectos psicológicos de los reactivos de las pruebas. Por ejemplo, Zeidner (1993) investigó las actitudes de los estudiantes hacia los formatos de los reactivos y descubrió que prefieren los reactivos de opción múltiple a los de ensayo. La investigación de Lu y Suen (1995) indica que la evaluación basada en el desempeño tiende a favorecer a los estudiantes independientes del campo sobre los dependientes (véase el capítulo 16). Otros investigadores han observado la relación entre la ansiedad ante la prueba y los reactivos, y han descubierto que la ansiedad parece afectar más a las pruebas de construcción que a las de selección de respuesta (Crocker y Schmitt, 1987). El lector puede encontrar en R. E. Snow (1993) un análisis de las variables que influyen en la motivación de los examinados y, por ende, afectan su respuesta y desempeño (por ejemplo, el propósito del procedimiento de evaluación), así como muchos otros factores que pueden tener relación con la interpretación de las pruebas de construcción de respuesta o de opción múltiple.

Al mismo tiempo, gradualmente se acumula la bibliografía empírica relativa a las propiedades estrictamente psicométricas de las tareas basadas en el desempeño que se utilizan en los medios académicos.[7] Tanto el ritmo de la investigación como la dirección de sus resultados difiere considerablemente dependiendo del tipo específico de reactivos en cuestión. En numerosos estudios se ha investigado la confiabilidad de los procedimientos de puntuación empleados para las tareas de construcción de respuesta, que (como los aplicados en las pruebas situacionales que vimos en el capítulo 16) a menudo consisten en valoraciones (E. L. Baker *et al.*, 1993; Linn y Gronlund, 1995, capítulo 10). Por lo general, la confiabilidad entre calificadores suele ser favorable cuando las reglas de calificación son claras y están bien desarrolladas y los calificadores capacitados. Por otro lado, la generalización de los resultados entre temas y tareas suele ser baja, lo que sugiere que los reactivos de construcción de respuesta tienen un elevado grado de especificidad de la tarea, lo que no es sorprendente si consideramos que estos reactivos suelen ser más complejos y permiten una mayor gama de respuestas que los reactivos de selección de respuestas.

En relación con el tema de la validez, la base empírica es más bien limitada, al menos para las clases menos restringidas y más innovadoras de tareas de construcción de respuesta. Una de las cuestiones preliminares más cruciales es el grado en que los reactivos de construcción y de selección de respuesta miden rasgos o habilidades equivalentes. Aunque las evidencias sobre este tema no son grandes, la revisión que hizo Traub (1993) de los estudios sobre el tema indican que el grado de equivalencia varía según el dominio; por ejemplo, cuando se utilizan diferentes formatos de reactivo en pruebas de comprensión de lectura o de conocimiento en matemáticas, tienden a arrojar resultados equivalentes, mientras que en el área escrita parece que la clase de los reactivos tiene un efecto significativo en las puntuaciones.

Mientras tanto, el uso de reactivos de opción múltiple sigue siendo considerable; de hecho, la introducción de diferentes formatos en los exámenes académicos aunada a las severas críticas hechas a los reactivos de opción múltiple parecen haber estimulado los esfuerzos por mejorarlos. Se dispone de orientación para desarrollar, revisar y evaluar los reactivos de opción múltiple (véase, por ejemplo, Haladyna, 1994) y con-

[7] Véase, especialmente, R. E. Bennett y Ward (1993).

tinúa la investigación sobre aspectos específicos (como el número óptimo de opciones presentadas) del formato (Trevisan, Sax y Michael, 1991, 1994). Más aún, también se estudian y difunden algunas variantes y perfeccionamientos del género de tareas de selección de respuesta (véase, por ejemplo, Linn y Gronlund, 1995, capítulo 8; Sax, 1991; Sireci, Thissen y Wainer, 1991; Wainer y Kiely, 1987; Wainer y Lewis, 1990).

También se han llevado a cabo comparaciones directas entre los reactivos de construcción y los de selección de respuesta (por ejemplo, Lukhele, Thissen y Wainer, 1994). La mayor parte de las veces, las comparaciones parecen favorecer a los reactivos de opción múltiple en criterios como economía, eficacia y validez predictiva, especialmente al compararse con las preguntas de ensayo tradicionales (Anastasi, 1988b, pp. 416-418; R. E. Bennett, 1993). Sin embargo, debe advertirse que el problema de la evaluación y comparación de diferentes formatos de evaluación (cuando las metas y los métodos de evaluación se encuentran en estado de cambio continuo) es cualquier cosa menos directo. Más aún, es importante tener en mente que el formato del reactivo es sólo una de muchas variables que interactúan para determinar la equidad, la exactitud y, sobre todo, la calidad de un procedimiento de evaluación. Es necesario considerar el propósito de la evaluación y el dominio de contenido así como las características de los individuos que deben ser evaluados (E. L. Baker *et al.*, 1993; R. E. Bennett, 1993; Dwyer, 1993; Mislevy, 1993). Por ejemplo, temas como los efectos diferenciales del fracaso en la motivación subsecuente del examinado pueden colocarse a la cabeza de otros criterios para evaluar los reactivos, en particular en el caso de los estudiantes que provienen de condiciones desfavorables o que presentan otras desventajas; no obstante, es digno de mención que en la actualidad no hay razón para creer que la evaluación basada en el desempeño da por resultado la disminución de la brecha entre las puntuaciones obtenidas por los blancos y los de algunas minorías étnicas en las pruebas estandarizadas de opción múltiple. En efecto, algunas investigaciones indican que esta disparidad en las puntuaciones puede ser mayor en los exámenes de construcción que en los de selección de respuesta (Hartle y Battaglia, 1993). Los temas actuales en la evaluación educativa de estudiantes con antecedentes culturales y lingüísticos distintos son analizados por Cancelli y Arena (1996), K. W. Howell y Rueda (1996), y por Shinn y Baker (1996).

Tipos de pruebas educativas

En el Informe Anual de 1990 del Servicio de Pruebas Educativas, el patronato de la organización predijo que la evaluación educativa cambiaría más en los 10 años siguientes que en los 50 anteriores (ETX, 1990). Parece ahora que la predicción fue exacta y que bien podría ser reiterada para la próxima década. Todos los exámenes están siendo sometidos a revisiones importantes y se preparan nuevas herramientas de evaluación a un paso rápido. En consecuencia, la siguiente revisión no se concentra en la descripción detallada de algunos exámenes individuales, sino en los instrumentos usados tradicionalmente en los contextos educativos, junto con algunas innovaciones en curso en cada uno.

Baterías de aprovechamiento general. Se dispone de algunas baterías para medir el aprovechamiento educativo general en las áreas que suelen cubrir los programas

académicos. Estas pruebas pueden utilizarse desde la primaria hasta el nivel adulto, aunque su principal aplicación ha sido en la escuela elemental. Por lo común, las baterías brindan perfiles de calificaciones en subtests individuales o en las principales áreas académicas. Una ventaja, en comparación con las pruebas de aprovechamiento construidas independientemente, es que permiten comparaciones horizontales, verticales o ambas. De este modo, es posible evaluar la posición relativa de un individuo en diferentes áreas temáticas o habilidades educativas en términos de una muestra normativa uniforme, o bien, informar el progreso del estudiante de un grado a otro en términos de una sola escala de calificación. El usuario de la prueba debe verificar que la batería está estandarizada de modo que permita uno o ambos tipos de comparabilidad.

Aunque algunas de las baterías de aprovechamiento general están diseñadas exclusivamente para los grados elementales o para la secundaria, la mayor parte abarca una gama amplia que se extiende en ambos niveles y a veces incluso hasta el primer año de la universidad. Algunas proporcionan una sola batería para el rango cubierto, pero casi todas tienen varias baterías sobrepuestas en cuadernillos de pruebas separados para usarse en diferentes niveles. Algunas baterías forman una serie coordinada, lo que permite la medición comparable del grado K al decimosegundo. Una de estas combinaciones es el Test de Habilidades Básicas de Iowa (*Iowa Tests of Basic Skills*), los Tests de Aprovechamiento y Destreza (*Tests of Achievement and Profiency*) y los Tests de Desarrollo Educativo de Iowa (*Stanford Achievement Tests Series*); otra es la Serie de Pruebas de Aprovechamiento de Stanford.

Una característica digna de mención de algunas baterías de aprovechamiento es que se normaron al mismo tiempo con los tests de inteligencia académica o de habilidad escolar. Los principales ejemplos incluyen las baterías de aprovechamiento que se parearon con tres pruebas de niveles múltiples ilustradas en el capítulo 10, a saber, la Serie de Pruebas de Aprovechamiento de Stanford con el Test de Habilidad Escolar de Otis-Lennon; la serie de Tests de Iowa y las Pruebas de Aprovechamiento y Destreza con el Test de Habilidades Cognoscitivas, y los Tests de Aprovechamiento de California y los Tests Comprensivos de Habilidades Básicas con el Test de Habilidades Cognoscitivas (véase la tabla 10.1). El uso de la misma muestra de estandarización en estos casos permite la comparación directa de las puntuaciones obtenidas por cualquier estudiante en las dos pruebas. Por lo general, ambas tienen una correlación elevada, y los individuos obtendrán puntuaciones muy similares en ellas. Para los estudiantes que obtienen una puntuación sustancialmente mayor en cualquiera, resulta deseable explorar las razones de la discrepancia. La batería de aprovechamiento mide en gran medida lo que el individuo ha aprendido en los cursos de la educación básica; los tests de habilidades cognoscitivas evalúan una gama más amplia de habilidades y conocimiento aprendidos dentro y fuera de la escuela. Cualquier discrepancia significativa en el desempeño podría reflejar la influencia de capacidades o discapacidades específicas o bien de factores no cognoscitivos como la motivación, los intereses y las actitudes. A menudo, los antecedentes del individuo ofrecen señales sobre las condiciones que dan lugar a discrepancias inusuales de ejecución en la prueba.

Por supuesto, las baterías de aprovechamiento difieren en el nivel técnico de sus procedimientos de construcción; no obstante, como grupo cumplen con estándares elevados, especialmente en relación con el tamaño y la representatividad de las muestras normativas, confiabilidad y validación de contenido. Las especificaciones detalladas

para la redacción de reactivos generalmente se siguen mediante cuidadosos análisis de reactivos, que incluyen la aplicación de los métodos de la teoría de respuesta al reactivo. Es común que se empleen procedimientos especiales para evitar los sesgos de género y étnico. Por lo general, todas las baterías incluyen la evaluación de habilidades básicas de lectura, lenguaje y matemáticas, en combinación con diversos montos del conocimiento contenido en los estudios científicos y sociales. Varias incluyen subtests diseñados para medir habilidades de estudio o la capacidad para emplear diversas fuentes de información. Más aún, en respuesta a las demandas de instrumentos de evaluación más flexibles e informativas, los editores de las principales baterías estandarizadas de aprovechamiento ahora ofrecen una mayor variedad de reactivos y opciones. Ahora se utilizan reactivos abiertos y una gama mayor de reactivos de selección de respuesta, diseñados para medir habilidades de pensamiento de orden superior y establecerse en contextos más significativos. Los editores muestran mayor flexibilidad para adaptar los paquetes de evaluación a las necesidades de los programas educativos actuales, mediante el uso de diversas combinaciones de formatos y contenidos de los reactivos así como de opciones de calificación. También ofrecen mayores conexiones entre las pruebas y los materiales educativos.[8]

Pruebas de competencia mínima en habilidades básicas. Las dos últimas décadas han atestiguado una preocupación creciente por el bajo nivel de competencia que muchos graduados de secundaria muestran en las habilidades de lectura, redacción y aritmética. Esta preocupación ha dado lugar a las demandas populares de las pruebas por competencia en habilidades básicas como una forma de certificar que se poseen las capacidades mínimas y como base para obtener el diploma de secundaria. Estas exigencias generaron una enorme controversia, y la mayor parte de las objeciones se concentró en los probables abusos e interpretaciones erróneas de tales pruebas y en la rigidez educativa y los controles burocráticos que pueden tener lugar.[9] Aunque la mayor parte de los estados ha establecido políticas concernientes a la evaluación de competencias mínimas, sus políticas y procedimientos varían ampliamente en el momento y el nivel en los que se aplican las pruebas, el uso que se hace de los resultados y la naturaleza y el grado de autonomía local en el desarrollo o elección de los instrumentos. Más todavía, las pruebas utilizadas para tomar decisiones sobre becas u obtención de diplomas deben reflejar el programa que cubren. Por todas estas razones, en la actualidad las pruebas apropiadas por lo general se elaboran en las escuelas, los sistemas escolares o las dependencias estatales, a menudo con la ayuda de editoriales especializadas, que pueden ofrecer personal capacitado, grandes bancos de reactivos y servicios de informes que pueden adaptarse para cumplir los objetivos locales. Estos instrumentos

[8] Un ejemplo de esa tendencia es la serie TerraNova, publicada en los Estados Unidos recientemente por CTB/Mc Graw-Hill. Los componentes de esta serie incluyen las nuevas Pruebas Comprensivas de Habilidades Básicas (*Comprehensive Tests of Basic Skills*, CTBS); una edición de Evaluaciones Múltiples, que combina reactivos de selección y de elaboración de respuestas; una edición de Evaluación del Desempeño, que ofrece tareas abiertas más extensas que pueden calificarse localmente o en la editorial, y un Componente de Costumbres, que ofrece reactivos complementarios diseñados para evaluar objetivos de programas especializados.

[9] Para un análisis general del movimiento de evaluación de competencias mínimas y los problemas psicométricos técnicos, véase Berk (1986). En el trabajo de Jaeger (1989) puede encontrarse otra revisión minuciosa de los temas y problemas del uso de las pruebas para certificar la competencia del estudiante.

pueden incluir algunos componentes de las baterías estandarizadas de aprovechamiento que vimos antes o pueden prepararse para cada localidad.

En los años recientes el interés por averiguar el dominio de las habilidades básicas se ha extendido a la población adulta.[10] Los efectos acumulativos de las grandes tasas de deserción escolar y los bajos niveles de competencia entre los graduados de secundaria, así como el aumento en el número de inmigrantes que no hablan inglés, han aumentado la preocupación por la competitividad de la fuerza de trabajo estadounidense en el mercado mundial. Los descubrimientos del Estudio Nacional de Analfabetismo de los Adultos (*National Adult Literacy Survey*), realizado en 1992 por el ETS bajo los auspicios del Departamento de Educación, han reforzado esta preocupación. El estudio reveló que casi la mitad de la población estadounidense se encontraba en los dos últimos niveles (de un total de cinco) de alfabetización (Kirsch, Jungeblut, Jenkins y Kolstad, 1993).

Las pruebas de competencia en habilidades básicas diseñadas para adultos suelen desarrollarse en relación con los cursos de educación para adultos, programas educativos en instituciones penales o programas de capacitación para el trabajo. Un ejemplo notable se encuentra en los Tests de Educación Básica para Adultos, formas siete y ocho (*Tests of Adult Basic Education*, TABE, 1994). La batería TABE abarca cinco niveles graduados de dificultad en cinco diferentes áreas de contenido que incluyen lectura, lenguaje y matemáticas aplicadas. Los resultados se asientan como calificaciones referidas a normas y también en términos de información basada en la competencia que puede utilizarse en forma de diagnóstico. Además de sus formas regulares, también puede tenerse acceso al TABE mediante una versión especial adecuada para emplearse en ambientes laborales y una edición en español (TABE Español) diseñada para medir las habilidades básicas de los adultos hispanohablantes en su idioma materno.

Pruebas para el salón de clases elaboradas por el maestro. Es indudable que los maestros preparan para usar en su salón de clases el mayor número de pruebas que cubren el contenido de cursos o de partes de cursos específicos. Es de todos conocida la gran diversidad que existe entre los cursos sobre el mismo tema y con el mismo nombre, especialmente a partir del nivel de secundaria, por lo que en esas condiciones ninguna prueba estandarizada externa puede bastar. Sin embargo, la preparación de pruebas para el aula puede mejorar con la aplicación de las técnicas y la experiencia acumulada de los profesionales en el desarrollo de pruebas. El desarrollo de pruebas para el aula puede dividirse en tres pasos principales: (1) la planeación del instrumento, (2) la redacción de los reactivos y (3) el análisis de los mismos. En el capítulo 7 explicamos algunas técnicas sencillas sobre análisis de reactivos adecuadas para usarse con grupos pequeños. En esta sección haremos una breve revisión general de los otros dos pasos.[11]

Es probable que un constructor de pruebas que se introduce directamente en la redacción de reactivos produzca una prueba asimétrica; si no se cuenta con un plan, al-

[10] Este análisis se fundamenta en el concepto de "analfabetismo funcional" (Sticht, 1975), que se ha extendido a la competencia promedio en el uso del lenguaje al hablar y escribir, así como a la comprensión y uso de diversos documentos y a cálculos aritméticos. La competencia funcional se define en términos de las demandas de las situaciones prácticas, como el nivel de dificultad y la cantidad de lectura requerida para realizar ciertos trabajos o, de manera más amplia, las habilidades educativas básicas exigidas para manejar la vida personal en la sociedad moderna.

[11] Para una guía adicional sobre la preparación de las pruebas en el salón de clases y otros procedimientos de evaluación, véase Linn y Gronlund (1995, capítulos 5 a 13).

gunas áreas quedarán sobrerrepresentadas mientras que otras prácticamente no serán tocadas. En general, es más sencillo preparar reactivos objetivos de algunos temas que de otros, y es más fácil preparar reactivos que requieran el recuerdo de hechos simples que otros que exigen una evaluación crítica, la integración de hechos diferentes o la aplicación de los principios a nuevas situaciones. De modo que si una prueba se elabora sin un programa de acción detallado, es probable que muestre una carga excesiva en el material inestable y menos importante. Muchas de las críticas a los instrumentos de selección de respuesta surgen del énfasis excesivo que los instrumentos mal construidos ponen en la memorización mecánica de detalles triviales.

Para prevenir esos desequilibrios fortuitos en la cobertura de los reactivos, es necesario establecer las *especificaciones de la prueba* antes de escribir cualquier reactivo. Para los exámenes en el salón de clases, estas especificaciones deberían basarse en el esbozo de los objetivos de instrucción y las áreas de contenido que deben cubrirse y deberían reflejar las decisiones sobre la relativa importancia de cada uno de estos aspectos mediante el número de reactivos asignado a cada tema y objetivo.[12] Quien desarrolla la prueba también debe decidir sobre la *forma de reactivo* más apropiada para el material; también es necesario considerar los méritos relativos de los reactivos objetivos y abiertos que vimos en una sección anterior de este capítulo en relación especial con su uso en los programas estandarizados de evaluación a gran escala. Por último, se han formulado muchas reglas prácticas para la *redacción de reactivos* sobre la base de años de experiencia en su preparación y en la evaluación empírica de las respuestas. Cualquiera que planee preparar un examen en el salón de clases haría bien en consultar alguna de las fuentes que resumen esas sugerencias (por ejemplo, Ebel, 1979, capítulos 4 a 9; Haladyna, 1994, capítulos 4 a 6; Linn y Gronlund, 1995, capítulos 6 a 9; Millman y Green, 1989).

Exámenes para nivel universitario. Se ha desarrollado una serie de pruebas y programas de evaluación para emplearlas en la admisión, colocación y consejería de los estudiantes universitarios. Uno de los más destacados es el Programa de Tests de Evaluación Escolar (*Scholastic Assessment Tests*, SAT) de la Junta Universitaria, que actualmente se compone de dos elementos, el SAT I: Prueba de Razonamiento, que reemplazó a las secciones verbal y matemática del Test de Aptitud Académica, y el SAT II: Pruebas Temáticas, que reemplazaron a las anteriores Pruebas de Aprovechamiento del SAT.[13] El SAT I consta principalmente de preguntas de opción múltiple que miden las habilidades verbal y matemática; está pensada para usarse (como complemento de las calificaciones de la educación media y otro tipo de información) en la evaluación de la preparación del estudiante para realizar el trabajo universitario. Por otro lado, las pruebas del SAT II están diseñadas para evaluar el conocimiento de áreas temáticas específicas (por ejemplo, literatura, química e historia universal) y puede emplearse lo mismo para admisión que para colocación.

Las pruebas del programa SAT han modificado tanto su contenido como su formato, lo mismo que el nombre; por ejemplo, en la sección verbal del SAT I hay un mayor

[12] En Anastasi (1988, p. 431) y en Linn y Gronlund (1996, p. 122) pueden encontrarse ejemplos de las tablas de especificación de tareas.

[13] Para una revisión adicional de los antecedentes del programa SAT, véase Anastasi (1988b, pp. 328–331) y Donlon (1984).

énfasis en la lectura crítica y el razonamiento; de igual modo, la sección de matemáticas requiere ahora que los estudiantes produzcan las respuestas (en lugar de limitarse a seleccionarlas) y (de acuerdo con los estándares actuales) se permite el uso de calculadoras en toda la sección de matemáticas. También han cambiado las pruebas del SAT II y en la actualidad incluyen una evaluación más directa de las habilidades mediante el uso de componentes del oyente y muestras de escritura.

A partir de abril de 1995, las calificaciones de las pruebas del programa SAT no se informan en términos de un grupo fijo de referencia que se remonta a 1941 (véase el capítulo 3). En lugar de ello, la escala de calificación del SAT se ha "recentrado" utilizando un nuevo grupo de referencia de la década de los noventa, de modo que la ejecución promedio se representa una vez más por una calificación aproximada de 500.[14] El recentramiento hizo más precisas y confiables las calificaciones del SAT, sobre todo en los niveles superior e inferior de la escala. Además, la interpretación de las calificaciones es más sencilla; por ejemplo, las calificaciones en las secciones verbal y de matemáticas ahora puede compararse directamente sin referirse a los rangos percentiles porque ambas han sido realineadas. Más aún, como una calificación de 500 es el punto medio del intervalo de 200 a 800, el promedio "intuitivo" sería igual al promedio real. La Junta Académica ha difundido tablas y otras herramientas que simplifican la conversión de las calificaciones de la escala original a las calificaciones de la escala recentrada para ayudar a mantener la continuidad entre ambas. Actualmente se dispone de información sobre la confiabilidad, niveles de dificultad y tasas de completamiento de los exámenes revisados del SAT en un manual preparado para consejeros y encargados de las oficinas de admisión así como de otras publicaciones e informes de investigación del ETS y de la Junta Universitaria (College Board, 1995a, 1995b).[15] En un estudio preliminar en el que se compara al SAT tradicional con una versión prototipo del SAT I se descubrió que el nuevo instrumento hacía una predicción ligeramente mejor que el anterior del promedio de calificaciones obtenido en el primer año en la universidad (Hale, Bridgeman, Lewis, Pollack y Wang, 1992). Evidencias adicionales sobre la validez de los exámenes revisados del SAT se incluirán en un complemento técnico cuya publicación se espera para finales de los noventa.

Otro programa nacional, iniciado en 1959, es el Programa Estadounidense de Evaluación Universitaria (*American College Testing Program*, ACT, 1995–1996), que aunque originalmente estaba limitado a los sistemas estatales de universidades, creció rápidamente y ahora se utiliza en muchas universidades de los Estados Unidos. La actual evaluación ACT incluye cuatro pruebas: de inglés, matemáticas, lectura y razonamiento científico. Este programa de examinación refleja el punto de vista de su fundador, E. F. Lindquist, y proporciona un conjunto de muestras de trabajo universitario. Se traslapa con los tests tradicionales de aptitud y de aprovechamiento, además

[14] Para principios de los noventa, las calificaciones promedio sobre la base del SAT habían disminuido de 500 en ambas áreas a 424 en el área verbal y a 478 en matemáticas.

[15] Debido a su función en la selección de estudiantes universitarios, los exámenes del SAT son sometidos con frecuencia a un escrutinio crítico; por ejemplo, hace poco se informó que los examinados podían responder correctamente muchas preguntas de opción múltiple basadas en las pruebas de lectura del SAT sin consultar los pasajes que acompañan a las preguntas, lo que ha reiniciado la controversia sobre el grado al que antecedentes extraños de conocimiento pueden influir en las calificaciones del SAT (véase, por ejemplo, S. Katz y Lautenschlager, 1995).

de que se concentra en las habilidades intelectuales básicas que se requieren para tener un desempeño satisfactorio en la universidad. Los componentes no cognoscitivos del programa ACT incluyen un cuestionario de información sobre las calificaciones obtenidas en secundaria, el inventario de intereses ACT y una Sección de Perfil del Estudiante que contiene preguntas sobre sus aspiraciones, planes y logros, así como otra información de sus antecedentes. El ACT no ha alcanzado los estándares técnicos establecidos por el SAT, pero los datos de validez se comparan favorablemente con los encontrados para otros instrumentos en ambientes similares.

Debe mencionarse que instrumentos como los del SAT y el ACT no pretenden sustituir las calificaciones de la educación media en la predicción del rendimiento en la universidad. Éstas pueden hacer una predicción de las calificaciones obtenidas en la universidad tan buena o ligeramente mejor que la mayor parte de las pruebas; sin embargo, cuando se combinan las calificaciones de las pruebas con las calificaciones escolares, mejora la predicción del rendimiento en la universidad. Esta mejora surge en parte del hecho de que una prueba uniforme y objetiva corrige la variabilidad de las normas de calificación de las diferentes secundarias. Además, las pruebas no están sometidas a los posibles sesgos personales u otros factores arbitrarios que pueden intervenir en la asignación de las calificaciones escolares.

También crece el uso de pruebas especializadas de aprovechamiento como equivalentes de los exámenes universitarios. Los estudiantes de educación media que tienen una preparación adicional en ciertas áreas pueden presentar examen en el Programa Avanzado de Colocación de la Junta Universitaria (*Advanced Placement Program*, APP) para obtener la admisión a la universidad y una posición avanzada en una o más materias. En el Programa de Exámenes de Nivel Universitario (*College Level Examination Program*, CLEP) se encuentra un desarrollo relacionado, cuyo propósito general consiste en facilitar el otorgamiento de créditos universitarios por el examen y ofrecer un sistema nacional para evaluar la educación de nivel universitario adquirida mediante estudio independiente y otros procedimientos no tradicionales. El Programa Estadounidense de Evaluación Universitaria aplica una serie similar de pruebas, el Programa de Exámenes de Destreza del ACT, que aunque incluye algunas materias académicas, como anatomía, fisiología y psicología anormal, cubre principalmente áreas ocupacionales como enfermería y contabilidad.

Admisión a la escuela de posgrado. La práctica de evaluar a los solicitantes de admisión se extiende a las escuelas profesionales y de posgrado. Casi todas las pruebas diseñadas con este propósito representan una combinación de tests de inteligencia general y pruebas de aprovechamiento. Un ejemplo conocido lo constituyen los Exámenes de Registro de Graduados (*Graduate Record Examination*, GRE), una serie que tuvo su origen en 1936 en un proyecto conjunto de la Fundación Carnegie para el Fomento de la Enseñanza y las escuelas de posgrado de cuatro universidades. Ahora que se ha expandido considerablemente, el programa es conducido por el Servicio de Evaluación Educativa bajo la dirección general del Consejo de Exámenes de Registro de Graduados. Los estudiantes son examinados en centros designados (en más de 100 países) antes de ser admitidos en las escuelas de posgrado. Los resultados son utilizados como auxiliares en la toma de decisiones sobre la admisión y colocación y para selec-

cionar a quienes recibirán becas, pensiones y designaciones especiales. El GRE incluye una Prueba General y Pruebas Temáticas en diversas áreas de especialización.[16] La Prueba General actual produce calificaciones independientes para las habilidades verbal, cuantitativa y analítica. En cuanto a las Pruebas Temáticas, se dispone de exámenes en 16 áreas que incluyen la biología, la computación, el francés, las matemáticas, la música, las ciencias políticas y la psicología. Las características psicométricas del GRE se informan en la última edición de la guía sobre su uso (*GRE 1995-1996 Guide*). En general, las calificaciones de las pruebas temáticas predicen mejor la calificación promedio obtenida en el primer año en la escuela de posgrado (*Grade Point Average*, GPA) que la calificación compuesta de la Prueba General o el GPA obtenido antes, pero la combinación de las tres medidas proporciona la mayor validez predictiva. En diversos campos, esas correlaciones múltiples caen entre la media de .40 e inferior de .60.

En octubre de 1992 el programa GRE empezó a aplicar una versión computarizada de la forma tradicional de la Prueba General, y en noviembre de 1993 se introdujo una Prueba General adaptada a la computadora. A pesar de algunos problemas iniciales relacionados con los riesgos de seguridad en la Prueba General computarizada del GRE, las ventajas del uso de computadoras es tal que es posible que para 1999 la Comisión del GRE elimine todos los exámenes de lápiz y papel. La Prueba General la están revisando para incluir una Prueba Escrita y una de Razonamiento Matemático, así como algunas preguntas de construcción de respuesta ("Update on the New GRE", 1995).

Evaluación de diagnóstico y pronóstico. A diferencia de las baterías de aprovechamiento general y otras pruebas que hemos estudiado, los instrumentos que veremos en esta sección están diseñados para analizar las fortalezas y debilidades de la persona dentro de un área temática particular y para sugerir las causas de sus problemas. La mayor parte de esos instrumentos de diagnóstico se aplica individualmente, por lo que suelen considerarse herramientas clínicas; sin embargo, algunos son componentes separados de las principales baterías de aprovechamiento que, según lo que ya vimos, están planeadas para la aplicación colectiva. Casi todas las pruebas colectivas de diagnóstico que se han publicado abordan habilidades de lectura, matemáticas y lenguaje y ofrecen información normativa y referida al contenido. Algunos ejemplos de la aproximación se encuentran en el Test de Diagnóstico en Matemáticas y el Test de Diagnóstico en lectura de Standford, así como en los Tests de Diagnóstico de Lectura y Matemáticas de California. Los editores de las dos series también han producido instrumentos separados para la evaluación y el diagnóstico de las habilidades de redacción. Tanto el Programa de Evaluación de Redacción de Standford como el sistema CTB de Evaluación de Redacción utilizan muestras directas de diversos modos de redacción —por ejemplo, narraciones o descripciones— y ofrecen diversas opciones de calificación.

En relación con el uso de todos los instrumentos de diagnóstico, hay un punto que merece especial énfasis. El diagnóstico de los problemas de aprendizaje y el subsecuente programa de la enseñanza de remedio son las funciones apropiadas del especialista capacitado. Ninguna batería de pruebas de diagnóstico es suficiente para este propósito.

[16] Antes de 1982, la Prueba General se denominaba Test de Aptitud y las Pruebas Temáticas se conocían como Tests Avanzados. Como sucedió con el SAT, los nombres cambiaron para evitar el riesgo de una mala interpretación sobre el propósito de las pruebas.

El diagnóstico y el tratamiento de los problemas graves de aprendizaje requieren un intensivo estudio de caso clínico, de preferencia interdisciplinario, que incluya la información complementaria sobre las capacidades sensoriales y el desarrollo motor, el historial médico y de salud, el historial educativo completo, datos sobre el hogar y los antecedentes familiares y una minuciosa investigación de las posibles dificultades emocionales. Aunque también las entrevistas y pruebas de diagnóstico colectivas pueden ayudar a identificar a los individuos que requieren mayor atención, el diagnóstico y el tratamiento de los problemas de aprendizaje requieren de técnicas especializadas. En el capítulo 9 estudiamos algunos de esos procedimientos, y volveremos al tema también en este capítulo en relación con la evaluación clínica.

Ciertas pruebas diseñadas para emplearse en los contextos educativos son en esencia instrumentos de pronóstico. Como tales, funcionan como tests de aptitudes más que como pruebas de aprovechamiento; al mismo tiempo, con frecuencia su contenido se parece al de estas últimas, ya que lo que toman como base para la predicción suele ser el rendimiento en un curso específico. El ejemplo más característico es el Test de Pronóstico de Álgebra de Orleans-Hanna (*Orleans-Hanna Algebra Prognosis Test* (Hanna, Sonnenschein y Lenke, 1983), en el que se entrega a los estudiantes un conjunto de "lecciones" breves de álgebra y se los examina de inmediato sobre lo que aprendieron. La prueba consiste entonces en muestras de trabajo en las que el aprendizaje que han de mostrar posteriormente en el curso se predice de su rendimiento en las tareas de muestra. Un ejemplo menos común y aún experimental de la evaluación de pronóstico se encuentra en las pruebas de lenguaje artificial desarrolladas por la Oficina de Manejo de Personal de los Estados Unidos y por el Departamento de la Defensa para predecir la habilidad para aprender un nuevo idioma (Diane, Brogan y McCauley, 1991).

Hay otro método de evaluación que, aun cuando tiene una connotación totalmente individualizada, ha recibido gran atención desde los años ochenta. En esencia, la aproximación sigue el procedimiento de examen-enseñanza-examen descrito como dinámica de evaluación guiada y se asocia con la instrucción de remedio. El potencial de aprendizaje del individuo se evalúa observando qué tan bien puede aprender en una relación individual con un profesional que cumple la triple función de examinador, instructor y clínico. Feuerstein (1979) es uno de los principales exponentes del método, y A. L. Brown, Campione, Webber y McGilly (1992) y Lidz (1987, 1997) analizan algunas aproximaciones relacionadas. Debido a su naturaleza clínica, consideraremos más a fondo la evaluación dinámica en una sección posterior del capítulo.

La medición basada en el currículo representa otro conjunto de técnicas elaborado para unir la evaluación con la intervención (Deno, 1992; L. S. Fuchs, 1993; Fuchs y Deno, 1991; Shinn, 1989). Aunque algunas aproximaciones de evaluación basadas en el currículo pueden ser muy informales, otras incluyen medidas estandarizadas de la ejecución del estudiante en habilidades básicas como la lectura, la ortografía y los cálculos aritméticos (para una comparación de los modelos véase Shinn, Rosenfield y Knutson, 1989). El común denominador de estas técnicas es una orientación fuertemente conductual y una relación directa con las tareas que componen el programa común con la educación elemental, en contraste con la aproximación inferencial y referida a normas de los instrumentos psicométricos tradicionales. La evaluación basada en el currículo se ha empleado principalmente en los ambientes de educación especial.

Evaluación en la educación de la niñez temprana. En las tres últimas décadas se han publicado muchos instrumentos nuevos para medir el desarrollo educativo de los niños pequeños, una actividad a cuyo monto y naturaleza han contribuido diversos factores (capítulos 9 y 12). La investigación sobre el desarrollo cognoscitivo temprano, el florecimiento de programas de educación preescolar y la difundida preocupación por los efectos de las desventajas culturales sobre la capacidad del niño para beneficiarse de la instrucción escolar han desempeñado funciones importantes.[17] Algunas de las pruebas fueron diseñadas principalmente para medir los resultados de la educación en la niñez temprana y funcionar como pruebas de aprovechamiento; otras se presentaron como instrumentos predictivos para evaluar la preparación del niño para la instrucción de primer grado. Sin embargo, los dos instrumentos se combinan de manera imperceptible y es común que cada uno cumpla cualquier propósito.

La *maduración para la escuela* se refiere en esencia a la obtención de las habilidades, los conocimientos, las actitudes, las motivaciones y otros rasgos conductuales apropiados que permiten al estudiante beneficiarse al máximo de la educación escolar. Estos requisitos son lo que J. McV. Hunt y Kirk (1974) denominaron las "habilidades de entrada" que el niño necesita para enfrentarse a la situación de enseñanza-aprendizaje que encontrará en el primer grado. La maduración se relaciona con ciertos niveles mínimos de desarrollo físico y sensoriomotor, obtenidos por maduración así como el aprendizaje previo. Cada vez es mayor el énfasis que se pone en el desarrollo jerárquico de conocimiento y habilidades, por medio del cual la adquisición de conceptos simples preparan al niño para el aprendizaje de conceptos más complejos en cualquier edad.

En general, las pruebas de maduración se aplican al entrar a la escuela. Aunque tienen mucho en común con los tests de inteligencia para los grados de primaria, destacan más las habilidades que son importantes para aprender a leer. También se presta alguna atención a los requisitos del pensamiento numérico y al control sensoriomotor requerido para aprender a escribir. Entre las funciones específicas que suelen cubrirse están la discriminación visual y auditiva, el control motor, la comprensión auditiva, el vocabulario, los conceptos cuantitativos y la información general. Una batería de maduración de gran uso es la sexta edición de las Pruebas de Maduración Metropolitana (*Metropolitan Readiness Tests*, MRT6; para una revisión de la edición anterior, véase Mabry, 1995, y Stoner, 1995). Una aproximación diferente a la maduración para la escuela está en los instrumentos que se concentran en la comprensión que tiene el niño de conceptos relacionales comunes, como el Test de Conceptos Básicos de Boehm, Revisado (*Boehm Test of Basic Concepts*) y la Escala de Conceptos Básicos de Bracken (*Bracken Basic Concept Scale*) (para revisiones, véanse Fitzmaurice y Witt, 1989; Linn, 1989; Turco, 1989; y Ysseldyke, 1989). El lector puede encontrar en Gredler (1992) un análisis del concepto de maduración para la escuela junto con los temas concernientes a su evaluación.

Comentarios finales. El campo de la educación se encuentra en un estado de cambio continuo que se ve reflejado en la evaluación educativa en todo el mundo, y

[17] El último de ellos culminó en la articulación de una meta educativa nacional que afirma que, para el año 2000, todos los niños de los Estados Unidos deberán empezar la escuela preparados para aprender (National Council on Education Standards and Testing, 1992).

es de esperar que el cambio persista y se intensifique en el futuro próximo.[18] Muchos observadores concuerdan en que es necesario integrar la evaluación y la instrucción de tal forma que estos aspectos de la empresa educativa se complementen para beneficio del estudiante[19] (H. Gardner, 1992; Nitko, 1989). La mayoría reconoce también que tanto la evaluación como la enseñanza pueden mejorar con los avances teóricos y empíricos. Ningún instrumento educativa o de evaluación bastará por sí solo, pues cada una tiene sus limitaciones; más aún, diferentes estudiantes requieren de aproximaciones distintas, por lo que la búsqueda de las mejoras debe continuar.

EVALUACIÓN OCUPACIONAL

Es costumbre que las pruebas psicológicas se empleen como apoyos en la toma de decisiones ocupacionales, que incluyen tanto la consejería individual como las decisiones institucionales concernientes a la selección y clasificación de personal. En esta sección revisaremos las aplicaciones de la examinación que atañen a la evaluación de la calificación ocupacional del individuo desde el punto de vista de las instituciones responsables de la selección, asignación y evaluación del personal.

Las organizaciones en los sectores empresarial e industrial, en los niveles federal, estatal y local de los gobiernos y en las distintas ramas de las fuerzas armadas utilizan casi todo tipo de prueba disponible en la toma de decisiones relacionadas con el personal. A menudo se han desarrollado baterías de aptitudes múltiples y tests de aptitudes especiales para cumplir propósitos ocupacionales, como las pruebas situacionales que explicamos en el capítulo 16. También se observa un uso creciente de los inventarios de personalidad (capítulo 13) y de los datos biográficos (capítulo 16). Landy y colaboradores ofrecen una breve revisión del uso de las pruebas y otros instrumentos en la selección y clasificación de personal; asimismo, es posible encontrar un tratamiento amplio del tema en tres libros recientes editados por Rumsey, Walker y Harris (1994), Schmitt, Borman *et al.* (1993) y Schuler, Farr y Smith (1993). Los principales aspectos del uso que se hace de las pruebas en la industria y las organizaciones se examinan de manera exhaustiva en varios capítulos del manual compilado por Dunnette y Hough (1990–1992).[20] La Sociedad para la Psicología Industrial y Organizacional (*Society for Industrial and Organizational Psychology*, SIOP, 1987) ha preparado y adoptado un conjunto de principios para la validación y el empleo de los procedimientos de selección de personal. A la vez que se interesan por la buena práctica en la elección, el desarrollo y la evaluación de todos los procedimientos de selección de personal, estos principios resultan sumamente pertinentes para las pruebas estandarizadas. Además,

[18] Véase, por ejemplo, el volumen de E. L. Baker y O'Neil (1994) en el que se evalúan las innovaciones tecnológicas en la educación y la capacitación, el análisis que R. E Snow y Lohman (1989) hacen de las implicaciones de la psicología cognoscitiva para la medición educativa y el trabajo de Oakland y Hambleton (1995) sobre los avances internacionales actuales en la evaluación académica.

[19] En efecto, se observa una rápida aparición de nuevos sistemas basados en computadoras que incorporan aspectos instruccionales, de evaluación y administración que pueden adaptarse individualmente. El programa de Aprendizaje Plus desarrollado por ETS para estudiantes adultos que necesitan mejorar sus habilidades académicas básicas y el programa SchoolVista de IBM para K-12 constituyen dos ejemplos del estado actual de estos instrumentos.

[20] El último volumen del manual, compilado por Triandis, Dunnette y Hough (1994), aborda temas de la psicología industrial y organizacional en diversas culturas de todo el mundo.

tanto los actuales *Estándares de la Examinación* (AERA, APA, NCME, 1985) como su revisión propuesta (véase el capítulo 1) tienen un capítulo sobre el tema de la evaluación en los medios laborales. Otra aplicación importante de la evaluación ocupacional, cubierta también en ambas versiones del mencionado trabajo, es la certificación y licencia de quienes se consideran calificados para la práctica de cualquiera de un número considerable de ocupaciones y profesiones.[21]

Como en el tratamiento que hicimos de las pruebas educativas, en esta sección nos concentraremos en los instrumentos diseñados especialmente para propósitos ocupacionales, además de los de mayor aplicación que vimos en otros capítulos. También examinaremos algunos de los procedimientos y temas incluidos en la elaboración y validación de pruebas en ambientes laborales.

Validación de las pruebas para el trabajo

Desde el punto de vista del empleado y del empleador, es obvia la importancia de que los individuos se coloquen en los trabajos para los cuales están calificados. La colocación adecuada también implica que los rasgos que sean irrelevantes para los requisitos del trabajo no afecten (favorable o desfavorablemente) las decisiones de selección. Si un test de habilidad mecánica requiere un nivel de comprensión de lectura muy superior al que exige el trabajo, su uso no conduce al mejor aprovechamiento del personal para ese puesto. Ya nos resulta familiar el simple hecho psicométrico de que es necesario averiguar la validez del instrumento para los usos particulares de la prueba, aunque ha adquirido nueva importancia por la preocupación generalizada acerca de la colocación laboral de minorías que sufren de desventajas culturales y educativas (véase el capítulo 18). Una prueba no válida o una que incluya elementos que no están relacionados con el trabajo considerado puede excluir de manera injusta a los miembros de grupos minoritarios que podrían realizar ese trabajo de manera satisfactoria.

Otra preocupación pertinente, tanto para las organizaciones como para la sociedad en su conjunto, surge de la relación demostrada entre la productividad en el trabajo y la validez de los instrumentos de selección (en el capítulo 6 revisamos los procedimientos para evaluar esta relación y los resultados comunes). Las ganancias y pérdidas estimadas en la productividad que se asocian con los aumentos y las disminuciones de la validez de los procedimientos de selección de personal son considerables; en las instituciones que contratan a muchos empleados, como las dependencias gubernamentales, el valor acumulado de estas ganancias y pérdidas es tan grande que merece una estrecha atención.

Durante muchas décadas la opinión prevaleciente en la psicología de personal era que las pruebas de selección deberían pasar por la validación de la escala completa contra criterios locales del desempeño en el trabajo (en los capítulos 5 y 6 estudiamos procedimientos concretos para esa validación de predicción de criterio); sin embargo,

[21] Anastasi (1988b, pp. 468–474) hace una revisión de la examinación en las profesiones, incluidos los exámenes aplicados a los psicólogos para la obtención de certificación y licencia. Para un examen más reciente de los temas relacionados con la validación y el uso de exámenes de licencia y certificación en general, véase el número especial sobre el tema de *Evaluation & the Health Professions*, compilado por LaDuca (1994).

en casi todas las situaciones resulta poco realista el estudio longitudinal de validación de la escala completa. Incluso en condiciones inusualmente favorables, en las que se tiene acceso a grandes muestras de empleados, se hacen evidentes varias limitaciones prácticas (véase, por ejemplo, Anastasi, 1972; J. T. Campbell, Crooks, Mahoney y Rock, 1973). En vista de los problemas prácticos para conducir una validación local de predicción de criterio de la escala completa, se han explorado diversos procedimientos alternativos.

Procedimientos globales para la evaluación del desempeño. Una aproximación a la selección de personal utiliza procedimientos de evaluación que se parecen, tanto como sea posible, a la situación laboral total, aunque este parecido nunca puede ser completo. El *puesto a prueba* es lo que más se acerca a una verdadera réplica del trabajo; pero incluso en ese caso, la brevedad del periodo y el conocimiento de que el empleo es a prueba puede influir en la conducta del trabajador de muy diversas maneras. Las *muestras de trabajo* representan otro intento por aproximarse a la verdadera ejecución laboral. Aquí, la tarea es realmente una parte del trabajo que debe ser realizado en el puesto, pero tanto la tarea como las condiciones de trabajo son iguales para todos los solicitantes. Algunas pruebas de muestras de trabajo se adaptan y corresponden a empleos particulares. La representatividad de la muestra de trabajo y la cercanía con que la tarea duplique las condiciones laborales son consideraciones esenciales; algunos ejemplos conocidos son los exámenes de manejo, así como las pruebas estandarizadas de habilidades para el trabajo de oficina como la mecanografía y la operación de diversas máquinas.

Algunas pruebas emplean la *simulación* para reproducir las funciones realizadas en el puesto, lo que se mezcla imperceptiblemente con las muestras de trabajo. Los ejemplos van de la operación de un punzón en miniatura a los simuladores para ingenieros en locomotoras y pilotos de aviones. Los simuladores han sido utilizados para propósitos de capacitación y evaluación en el programa espacial estadounidense de la NASA, así como en diversas especialidades militares.

A esta lista deben agregarse las *técnicas del centro de evaluación* (véase el capítulo 16) que han sido extensamente empleadas en la evaluación del personal de gerencia o administrativo (Bray, 1982; Finkle, 1983; Moses, 1985; Thornton y Byham, 1982). Una característica distintiva de esta aproximación es la inclusión de las pruebas situacionales, como la de la *cesta*, técnica adaptada para evaluar a los ejecutivos en muchos contextos (N. Frederiksen, 1962, 1966; Shapira y Dunbar, 1980). Al simular la familiar "cesta" que se encuentra en el escritorio del administrador, la prueba proporciona un conjunto cuidadosamente preparado de cartas recibidas, memoranda, informes, documentos por firmar y artículos similares. Antes de presentar la prueba, el examinado tiene la oportunidad de estudiar los materiales para recibir orientación e información relacionadas con el trabajo hipotético. Durante la prueba, la tarea consiste en manejar todos los asuntos en la cesta que el examinado haría en el trabajo; todas las acciones deben registrarse por escrito, pero pueden incluir cartas, memos, toma de decisiones, planes, directrices, información que debe obtenerse o transmitirse, agenda de reuniones o cualquier otra nota. Otras técnicas del centro de evaluación pueden emplear la representación de papeles, la resolución de problemas en grupo y los juegos

de negocios. Un aspecto común es el uso de asesores múltiples y de valoraciones de los pares; muchos de los rasgos evaluados son pertinentes a la motivación, las habilidades interpersonales y otras variables de personalidad.

Aunque dependen al menos en parte del parecido con el puesto como evidencia de que están relacionados con el mismo, estos procedimientos globales de evaluación del desempeño también han sido calificados, por separado o en diversas combinaciones, contra diversos criterios (véase Landy *et al.*, 1994; Schmidt, Ones y Hunter, 1992).

Análisis de puestos y el método de elementos del trabajo. Hay un interés creciente en la aplicación de la validación de contenido a las pruebas de selección de personal. En todas sus formas, esta validación depende de un análisis minucioso y sistemático del puesto (McCormick, 1979). Para ser eficaz, el análisis del puesto debe identificar los requisitos que lo distinguen de otros, por lo que para este propósito resulta inútil una descripción en términos de generalidades que podrían aplicarse igualmente a la mayor parte de los puestos. Para obtener una imagen completa de las actividades del puesto, el analista debe recurrir a distintas fuentes de información. A este respecto, es común que se consulten manuales de entrenamiento y operación, registros de desempeño y, particularmente, a expertos en la materia como supervisores, instructores o trabajadores experimentados en el campo.

Un análisis eficaz del puesto también debe concentrarse en los aspectos del desempeño que hacen una buena diferenciación entre los mejores y los peores empleados. En su libro clásico, *Aptitude Testing*, Hull (1928) subrayó la importancia de estos aspectos que permiten hacer una diferenciación en el desempeño del trabajo. Posteriormente, el concepto lo subrayó una vez más J. C. Flanagan (1949, 1954), quien propuso una técnica de incidentes críticos que en esencia requiere descripciones reales de casos de conducta laboral que son característicos de los buenos o los malos trabajadores.

La atención en los requisitos críticos del puesto llevó al establecimiento del método de elementos del trabajo para construir pruebas y demostrar su validez de contenido (McCormick, 1979, 1983; McCormick, Jeanneret y Mecham, 1972; Primoff, 1975; Primoff y Eyde, 1988). Dicho método fue desarrollado y empleado por Primoff y sus colaboradores en la Oficina de Manejo de Personal de los Estados Unidos (antes conocida como la Comisión de Servicio Civil de los Estados Unidos). En esencia, los elementos del trabajo son las unidades que describen los requisitos cruciales del puesto. Aunque las diversas adaptaciones del método difieren en los detalles del procedimiento, todas proporcionan la descripción de las actividades laborales en términos de los requisitos conductuales específicos a partir de los cuales pueden formularse directamente los reactivos de la prueba. Las afirmaciones conductuales específicas a su vez pueden agruparse en categorías más amplias, o constructos, como la exactitud del cálculo, la destreza de manos y brazos, la discriminación visual o la capacidad para trabajar bajo presión. Existe una gran cantidad de investigación orientada al desarrollo de una taxonomía general del desempeño en el trabajo en términos de constructos conductuales amplios (Fleishman, 1975; Fleishman y Quaintance, 1984; Fleishman y Reilly, 1992b).

Los métodos de análisis de puestos pueden contribuir a la facilitación del uso eficaz de pruebas para muchos trabajos aparentemente diferentes, lo que ilustran instrumentos como el Estudio de Análisis de Puestos de Fleishman (*Fleishman Job Analysis*

Survey, F-JAS) y el sistema de Claves de Trabajo. El F-JAS es un instrumento de análisis de puestos diseñada para describir al puesto en términos de los conocimientos, las habilidades y las destrezas requeridas para realizarlo. Cincuenta y dos de sus 72 escalas cubren habilidades cuidadosamente definidas en los dominios cognoscitivo, psicomotor, físico y sensorial-perceptual, y casi todas han sido asociadas con los instrumentos existentes (Fleishman y Mumford, 1991; Fleishman y Reilly, 1992a, 1992b). Las 20 escalas restantes, que aún se encuentran en desarrollo, abordan los dominios interpersonal/social y de conocimiento/habilidades. Por otro lado, el sistema de Claves de Trabajo —formulado recientemente por el Programa de Evaluación de las Universidades Estadounidenses (ACT, 1995; Scruggs, 1994)— se concentra en un conjunto mucho menor de habilidades genéricas en los lugares de trabajo, como "localizar información" que puede enseñarse en un periodo razonable; sin embargo, en este marco de trabajo el sistema proporciona un paquete coordinado de instrumentos para: (1) hacer el análisis y el perfil del puesto; (2) evaluar los niveles de habilidad; (3) dar retroalimentación a los individuos, educadores y empleados, y (4) proporcionar apoyo educativo al poner en práctica el programa educativo o de capacitación.

El análisis de puestos es uno de los métodos más antiguos y viables de la psicología industrial. Sus aplicaciones en la validación de las pruebas de empleo han seguido aumentando en la medida en que los avances en la tecnología de la computación han hecho más económica la obtención y el análisis de datos laborales.[22] Además, la información obtenida mediante el análisis cuidadoso del puesto puede tener muchos otros usos, como establecer el valor de un trabajo en el mercado laboral o diseñar puestos (véase, por ejemplo, Campion, 1994; I. L. Goldstein, Zedeck y Schneider, 1993).

La predicción del desempeño laboral. Las dificultades prácticas inherentes a la conducción de la validación de predicción de criterios locales (capítulos 5 y 6) han dado lugar a una relativa escasez de tales estudios. Aun así, permanece el hecho de que muchas organizaciones requieren previsiones del futuro desempeño en el puesto para tomar decisiones relativas a la selección y colocación del personal.[23] Con tales propósitos, las técnicas de validación sintética y la generalización de la validez se han convertido en las opciones a elegir. Ambas permiten hacer estimaciones de la validez del instrumento para determinado trabajo en ausencia de la validación local y, en la medida en que se acumulan datos empíricos, los dos métodos deberían proporcionar evidencias convergentes sobre la naturaleza de los constructos evaluados (J. P. Campbell, 1990a).

El concepto de *validación sintética* se basa en la premisa del método de elementos del trabajo de que es posible identificar las habilidades, conocimiento y otros requisitos de ejecución que son comunes a muchos empleos diferentes. La validez sintética ha sido definida como "la inferencia de validez en una situación específica a partir del análisis sistemático de los elementos del trabajo, la determinación de la validez de la

[22] El lector puede encontrar en Harvey (1991) una reseña general de los avances y los problemas en la metodología del análisis de puestos; por su parte, Knapp, Russell y Campbell (1993) describen aplicaciones específicas del análisis de puestos en el contexto de la selección y clasificación del personal de las fuerzas armadas.

[23] El lector puede encontrar en De Wolff (1993) un acalorado debate sobre las limitaciones de la predicción como paradigma en la selección de personal.

prueba para esos elementos, y la combinación de las distintas instancias elementales de validez en un todo" (Balma, 1959, p. 395). En esencia, la técnica incluye tres pasos: (1) un análisis detallado del puesto para identificar sus elementos y el peso relativo de éstos en un trabajo particular; (2) el análisis y estudio empírico de cada prueba para determinar el grado en que miden la eficiencia en el desempeño de cada uno de esos elementos del trabajo, y (3) encontrar sintéticamente la validez de cada prueba para el trabajo dado a partir de los pesos de esos elementos en el puesto y en la prueba. Primoff (1959; Primoff y Eyde, 1988) elaboró un procedimiento estadístico para calcular esta validez. Designado como coeficiente J (por "*Job coefficient*" coeficiente del trabajo), el procedimiento es en esencia una adaptación de las ecuaciones de regresión múltiple que estudiamos en el capítulo 6. Hollenbeck y Whitemer (1988) y Mossholder y Arvey (1984) han descrito otras aproximaciones a la validación sintética.

Los procedimientos de generalización de la validez —desarrollados primero por Schmidt y Hunter (1977) y que vimos en el capítulo 5— proporcionan otro medio para validar las pruebas de selección de personal. Básicamente, la aproximación permite aplicar los hallazgos de validez a nuevas situaciones mediante las técnicas de meta-análisis (Schmidt, Hunter, Pearlman y Hirsh, 1985). Algunos han cuestionado el grado al que pueden generalizarse los datos obtenidos mediante meta-análisis; los críticos han subrayado las diferencias entre las situaciones de empleo así como los problemas metodológicos en la estimación de los parámetros. A su vez, esto ha dado lugar a perfeccionamientos en los métodos meta-analíticos y, por ende, a su mayor aceptación y aplicación. Aunque persisten algunos puntos polémicos, y aún hay posibilidad de mejoras, es innegable que los métodos para la generalización de la validez han contribuido considerablemente a la vitalidad de la teoría, investigación y práctica de la evaluación ocupacional (véase, por ejemplo, L. R. James *et al.*, 1992; Landy *et al.*, 1994; Schmidt *et al.*, 1993).

El criterio del rendimiento en el trabajo. Algunos de los trabajos más prometedores en el campo de la selección y clasificación de personal provienen de la renovada atención que se está brindando a los criterios. Se recordará de lo que vimos en el capítulo 5 que existe un vasto arreglo de indicadores que, dependiendo de cómo se defina al criterio, pueden considerarse como medidas del mismo. No obstante, en cada estudio de validación suele emplearse una única medida conveniente del desempeño laboral para representar "el" criterio, independientemente del propósito del proceso de predicción. Hasta hace muy poco tiempo, y a pesar de que durante varias décadas se exigió reiteradamente una consideración más cuidadosa de los criterios (véase, por ejemplo, L. R. James, 1973; Tenopyr, 1986; Wallace, 1965), fue poco lo que se hizo acerca de este problema crucial; sin embargo, en los últimos años varios investigadores han estado trabajando para lograr una conceptuación más clara del desempeño laboral y una mejor comprensión de sus determinantes (Borman, 1991; Campbell, McCloy, Oppler y Sager, 1993; B. F. Green y Wigdor, 1991; Schmidt y Hunter, 1992).

Un nuevo modelo del desempeño laboral que promete tener una importante función heurística es la teoría de factores múltiples debida a John P. Campbell y sus colaboradores en conjunto con el Proyecto de Selección y Clasificación del Ejército Estadounidense (Proyecto A; J. P. Campbell, 1990a, 1990b, 1994; Campbell, McHenry

y Wise, 1990). El modelo de Campbell tiene en consideración la naturaleza multifacética del desempeño laboral y separa los diversos elementos incluidos bajo el rubro. Al principio, el modelo hace algunas distinciones fundamentales entre los aspectos de la evaluación del trabajo que están bajo el control del trabajador (por ejemplo, las conductas comprometidas en el propio desempeño laboral) y los que no lo están —digamos, las consecuencias del desempeño laboral (eficacia), sus costos relativos (productividad) y el valor que la empresa asigna a cada uno de estos aspectos por la organización (utilidad)—. En lo que respecta al propio desempeño laboral, la teoría postula que cualquier empleo se vincula a múltiples componentes del desempeño (tareas) y que los determinantes de cada componente consisten en diversas combinaciones de elementos de conocimiento, habilidad y motivación en el trabajador. Más aún, cada determinante del desempeño laboral tiene ciertos antecedentes más o menos especificables (como la capacitación, las contingencias de reforzamiento y algunas características individuales) que pueden afectar el desempeño de manera indirecta por sus efectos sobre el nivel de conocimiento, habilidades y motivación del individuo. Además, estos determinantes del desempeño laboral interactúan, con un impacto consecuente sobre la ejecución.

Aunque la teoría de factores múltiples del desempeño laboral aún está en evolución (J. P. Campbell, 1990a, 1994; D. J. Knapp y Campbell, 1993), su diseño es bastante compatible con otros importantes avances conceptuales y metodológicos en la evaluación de la conducta laboral (Borman, 1991). Actualmente, el modelo identifica ocho factores generales del desempeño laboral que incluyen características como la consistencia del esfuerzo, la disciplina personal, el liderazgo, la eficiencia en las tareas específicas del trabajo y otras clases de destrezas. Se supone que estos factores son lo bastante amplios para abarcar los principales elementos necesarios para describir todos los empleos citados en el *Dictionary of Occupational Titles*. Además, el modelo especifica tres clases de determinantes de las diferencias individuales en el desempeño laboral, a saber, conocimiento declarativo, conocimiento de procedimientos y habilidad, y motivación, así como sus antecedentes. Esta estructura teórica amplia y claramente estructurada debería demostrar ser aplicable a una amplia gama de investigación realizada sobre los constructos del desempeño laboral.

Uso ocupacional de las pruebas

Como sugiere nuestra revisión de la sección anterior, lo adecuado del uso de las pruebas en las decisiones concernientes al personal no puede considerarse de manera desligada de los propósitos, las situaciones y las poblaciones incluidas en un determinado contexto.[24] También debe advertirse que, aun cuando para fines de la revisión podamos categorizar a los instrumentos por tipos, en la práctica no siempre resulta clara la línea de demarcación entre conocimientos, capacidades, habilidades y rasgos de personalidad, por lo que puede ser más fructífero pensar que, como algunos han planteado (Lubinski y Dawis, 1992), la conducta está determinada por las *capacidades de respuesta*. Más aún, aunque a menudo se revisa por separado la validez de las pruebas,

[24] Véase Sternberg (1994a) para una revisión breve de la multiplicidad de las variables que interfieren en el éxito de las interacciones entre la persona y el contexto.

dichos instrumentos apenas se utilizan como tales. La mayor parte de las decisiones relativas al personal basadas en las pruebas emplean una combinación de una o más mediciones además de otros instrumentos de evaluación, como entrevistas o los datos antecedentes.[25] Hechas estas advertencias, pasemos ahora a considerar el uso de las pruebas en los medios ocupacionales.

La función de la inteligencia académica. "Inteligencia" es un término amplio del que existen muchas definiciones; es indudable que lo que constituye la inteligencia varía según las culturas, las diferentes épocas y las etapas de la vida (capítulos 11 y 12). En contraste, los tests tradicionales de inteligencia cubren un grupo más restringido e identificable de conocimientos y habilidades cognoscitivas que, no obstante, ha demostrado ser ampliamente predictivo del desempeño en las actividades académicas y ocupacionales requeridas en las sociedades tecnológicas modernas. Debido a que trata ante todo con los conocimientos y las habilidades que se adquieren en el curso de la escolaridad formal en tales sociedades, a menudo se describe este agrupamiento de habilidad como inteligencia académica o habilidad escolar. Su contenido incluye en particular comprensión verbal, razonamiento cuantitativo y otros aspectos del pensamiento abstracto.

Es bien sabido que la ejecución en los tests de inteligencia académica correlaciona sustancialmente con el nivel de escolaridad, por lo que parecería posible establecer los requisitos educativos necesarios para cubrir la calificación del solicitante en este importante grupo de conocimientos y habilidades cognoscitivas; sin embargo, existen dificultades en la forma de esta solución. El nivel de escolaridad es un indicador indirecto del estado de desarrollo cognoscitivo del individuo, y la correlación entre ambos está lejos de ser perfecta. La simple exposición a la escolaridad formal no garantiza el mismo aprendizaje de lo que se enseña. Más aún, *es posible* adquirir por otros medios los conocimientos y las habilidades que se reciben en la escuela, por lo que es más justo para el individuo examinar sus conocimientos y sus habilidades cognoscitivas en lugar de aceptar o rechazar a los solicitantes sobre la base de la cantidad de la educación formal recibida.

Entre los instrumentos disponibles, se han desarrollado varios tests cortos de inteligencia académica para uso especial en la industria. Un ejemplo es el Test de Personal de Wonderlic (*Wonderlic Personnel Test, Inc.*, 1992), instrumento que empezó como una revisión de un grupo inicial de tests de inteligencia (el Test Autoaplicado de Habilidad Mental de Otis). Consta de 50 reactivos y se realiza en 12 minutos; incluye una variedad de reactivos con contenido verbal, numérico y espacial —presentado en un formato de espiral colectiva— y produce una sola calificación. Esta prueba, que se encuentra disponible en varias formas, durante el tiempo en el que ha estado en uso ha acumulado normas extensas sobre diversos grupos ocupacionales y su valor predictivo para el éxito en la capacitación y en el trabajo está bien documentado (para revisiones véase Belcher, 1992; Schmidt, 1985; y Schoenfeldt, 1985).

El interés en la utilidad potencial de las pruebas de inteligencia académica general para la selección de personal ha sido reavivado por la investigación sobre la generalización de la validez (véase, por ejemplo, Hunter, 1986). De particular relevancia son

[25] El lector puede encontrar en Guion (1991) una excelente revisión del proceso de evaluación, selección y colocación del personal que incluye los descubrimientos concernientes a la validez y la equidad de diversas pruebas y otros predictores.

los hallazgos de que las pruebas de razonamiento verbal y numérico tienen cierta validez predictiva para una amplia variedad y gama de trabajos. Además, la validez aumenta para aquellos trabajos con mayores exigencias de toma de decisiones y procesamiento de información; sin embargo, aunque las pruebas de habilidad cognoscitiva general contribuyen sustancialmente a la predicción del desempeño en el trabajo, en particular en los puestos complejos, la exactitud de la predicción puede aumentar mediante la evaluación de variables adicionales, entre las cuales las más importantes son las habilidades y los conocimientos especializados que requieren ciertos puestos —incluyendo las habilidades psicomotoras así como los conocimientos tácitos o de procedimiento—, y también variables no cognoscitivas como las características de temperamento y actitudes (véase, por ejemplo, Ackerman, 1992; J. P. Campbell, 1990b; Carroll, 1992; Kanfer, Ackerman, Murtha y Goff, 1995; Sternberg, Wagner, Williams y Horvath, 1995). Muchas de estas relaciones se han demostrado en proyectos de investigación a gran escala con las baterías de clasificación empleadas por las fuerzas armadas y ciertas dependencias gubernamentales.

Baterías de aptitudes para programas especiales.[26] El Servicio de Empleo de los Estados Unidos (*United States Employment Service*, USES) elaboró la Batería de Pruebas de Aptitudes Generales (*General Aptitude Test Battery*, GATB) para el uso de los consejeros laborales en las oficinas estatales de servicio de empleo (U. S. Department of Labor, 1970). Además, la batería puede ser obtenida por organizaciones no lucrativas como escuelas secundarias, universidades y prisiones. En la actualidad, la GATB comprende 12 pruebas, cuatro de las cuales requieren aparatos sencillos mientras que las ocho restantes son de lápiz y papel. Puede aplicarse aproximadamente en dos horas y media, arroja calificaciones en nueve factores y en tres medidas compuestas principales derivadas de los factores, los cuales se presentan en la tabla 17.1.

La utilización de las calificaciones de la GATB puede hacerse mediante dos aproximaciones distintas. La primera hace uso de las puntuaciones de corte múltiples en las aptitudes más importantes requeridas por grupos relativamente homogéneos de trabajos. Un mecanismo usado con esta aproximación es la estructura del Patrón de Aptitudes Ocupacionales (*Occupational Aptitude Pattern*, OAP), definida en los años

Tabla 17.1

Batería de Pruebas de Aptitudes Generales (GATB) factores y compuestos

Factores		
G. *Capacidad gral. de aprendizaje*	S. *Aptitud espacial*	K. *Coordinación motora*
V. *Aptitud verbal*	P. *Percepción de formas*	F. *Destreza de los dedos*
N. *Aptitud numérica*	Q. *Percep. del trabajo de oficina*	M. *Destreza manual*
Compuestos		
Cognoscitivo = G + V + N	**Perceptual** = S + P + Q	**Psicomotor** = K + F + M

[26] En los capítulos 10 y 11 se pueden encontrar los antecedentes relacionados con esta sección.

setenta (U. S. Department of Labor, 1979, 1980). Se han preparado patrones para más de 60 familias de trabajos que cubren miles de ocupaciones. Para cada grupo de trabajo, se han calculado las puntuaciones de corte que denotan niveles alto, medio y bajo de calificación en las aptitudes pertinentes, las que pueden emplearse para aconsejar a los individuos respecto a las carreras más adecuadas.[27] Otra aproximación al uso de la GATB se derivó de la aplicación de técnicas de generalización de la validez (GV) a los datos de más de 500 estudios de validez realizados por el Servicio de Empleo de los Estados Unidos. Este procedimiento, que llegó a conocerse como GV-GATB, utiliza estimaciones de validez basadas en las combinaciones apropiadas de puntuaciones para todos los trabajos incluidos en cada familia (U. S. Department of Labor, 1983a, 1983c, 1983d). Las predicciones se basan en tres medidas compuestas (cognoscitiva, perceptual y psicomotora) derivadas de las puntuaciones de los factores originales. De los tres, el compuesto cognoscitivo arroja los coeficientes de validez más elevados para casi todos los empleos, pero el compuesto psicomotor puede mejorar la predicción en la medida en que disminuye la complejidad del puesto (Hunter y Hunter, 1984).

Gracias al USES se ha acumulado un extenso cuerpo de datos sobre la GATB, gran parte del cual atestigua su excepcional confiabilidad y considerable validez predictiva (para revisiones, véase B. Bolton, 1994; Kirnan y Geisinger, 1986). Sin embargo, la práctica de establecer normas para subgrupos (que en la década de los ochenta se estableció para la GV-GATB como una forma de asegurar valores de referencia comparables para solicitantes blancos, negros e hispanos) dio lugar a un acalorado debate acerca de la equidad de la examinación previa al empleo (Hartigan y Wigdor, 1989), el debate culminó en la aprobación de la Ley de Derechos Civiles de 1991 (P. L. 102–166) que prohíbe las normas para subgrupos. La legislación afectó el uso de la batería e hizo incierto su futuro (L. S. Gottfredson, 1994; Wigdor y Sackett, 1993, véase también el capítulo 18). No obstante, todavía está en marcha un programa de investigación que incluye el desarrollo de dos nuevas formas y una versión adaptada a la computadora experimental de la GATB.

Otro instrumento importante de selección y clasificación es la Batería de Aptitud Vocacional de las Fuerzas Armadas (*Armed Services Vocational Aptitude Battery*, ASVAB), desarrollada para usarse en todas las ramas militares de los Estados Unidos (Bayroff y Fuchs, 1970). Se aplica a los estudiantes de educación media interesados en las ocupaciones militares y a los individuos que han solicitado su ingreso a la milicia. Las formas actuales de la ASVAB incluyen 10 subtests que se presentan en la tabla 17.2.[28] El Test de Calificación de las Fuerzas Armadas (*Armed Forces Qualification Test*, AFQT) es una puntuación compuesta utilizada por todos los servicios como indicador de la posibilidad general de capacitación en la selección de los reclutas. Además, cada

[27] Los grupos de la Batería de Pruebas de Aptitudes Especiales (*Special Aptitude Test Battery*, SATB) proporcionan un mecanismo alternativo para el uso de las puntuaciones de corte con la GATB. Estudiamos más a fondo la estrategia de cortes múltiples en el capítulo 6. En el campo de la selección, a diferencia de la consejería, el uso más apropiado de las puntuaciones de corte consiste en la identificación preliminar de los solicitantes en una o más habilidades críticas.

[28] Se cuenta también con una versión de Evaluación Adaptada por Computadora de la ASVAB (CAT-ASVAB), que ha estado en desarrollo desde 1979 y que ahora se usa operacionalmente para algunos procedimientos de ingreso militar (T. L. Russell, Reynolds y Campbell, 1994). Véase Wiskoff y Schratz (1989) para una descripción del CAT-ASVAB y de su desarrollo.

Tabla 17.2

Batería de Aptitudes Vocacionales de las Fuerzas Armadas (ASVAB)

Subtests ASVAB	
Ciencia general (GS) [General Science]	Razonamiento aritmético (AR)[a] [Arithmetic Reasoning]
Conocimiento del mundo (WK)[a] [Word Knowledge]	Conocimiento matemático (MK)[a] [Mathematics Knowledge]
Comprensión de párrafos (PC)[a] [Paragraph Comprehension]	Comprensión mecánica (MC) [Mechanical Comprehension]
Información electrónica (EI) [Electronics Information]	Información de autos y talleres (AS) [Auto and Shop Information]
Rapidez de codificación (CS)[b] [Coding Speed]	Operaciones numéricas (NO)[b] [Numerical Operations]

[a]Compuesto AFQT [b]Velocidad

uno de los servicios combina los subtests para formar compuestos que correspondan a sus propias necesidades de selección y clasificación de personal; por ejemplo, el compuesto de Combate de la Armada consta de los componentes de los subtests AR + CS + AS + MC. Las calificaciones estándares para la ASVAB se basan en las normas obtenidas de una muestra representativa de jóvenes estadounidenses (U. S. Department of Defense, 1982). La estructura factorial de la batería ha sido examinada con cierta profundidad, y los resultados por lo común muestran un factor general que da cuenta de alrededor del 60 por ciento de la varianza total de la ASVAB y de cuatro factores grupales que han sido replicados repetidamente (J. R. Welsh, Watson y Ree, 1990). Los cuatro factores, junto con los subtests en que tienen las mayores cargas son: (1) Verbal (WK y PC); (2) Velocidad (NO y CS); (3) Cuantitativo (AR y MK), y (4) Técnica (AS, MC y EI).

Se ha investigado la validez de los subtests y los compuestos de la ASVAB en comparación una gran variedad de criterios de desempeño laboral y educativo. Como era de esperarse, los coeficientes de validez difieren sustancialmente según el tipo y la cantidad de criterios empleados. En general, la validez es mayor para los criterios de "puede hacer" (como el servicio militar en general y la eficiencia técnica) que los criterios de "hará" (como un esfuerzo, liderazgo y disciplina personal). Como se suponía, los primeros se evalúan mediante las calificaciones asignadas por el supervisor, los pares y por el propio individuo (McHenry, Hough, Toquam, Hanson y Ashworth, 1990; T. L. Russell *et al.*, 1994).

En 1980, el Departamento de Defensa de los Estados Unidos emprendió el examen más extenso de la ASVAB como parte del Proyecto del Servicio Conjunto de Medición del Desempeño Laboral y de los Estándares de Reclutamiento (*Job Performance Measurement*, JPM) (Wigdor y Green, 1991a, 1991b). Este proyecto masivo pretende imponer medidas sólidas de desempeño para el nivel de ingreso a trabajos militares de forma que sea posible establecer estándares significativos y válidos de reclutamiento para todos los voluntarios que solicitan su ingreso a las fuerzas armadas. La primera fase

del JPM estableció que la ASVAB es un buen predictor de alta fidelidad, indicadores basados en el desempeño de la eficiencia en el trabajo.[29] También proporciona apoyo a la validez diferencial de las puntuaciones compuestas de la ASVAB para diferentes puestos; sin embargo, la magnitud de las diferencias en la puntuación promedio entre examinados negros y de grupos no minoritarios fue considerablemente mayor en la AFQT y en las pruebas de lápiz y papel del conocimiento laboral que en las muestras de prueba de ejecución en el trabajo, lo que indica la posibilidad de que algunas de las medidas ASVAB puedan sobrevalorar la magnitud de las verdaderas diferencias grupales en el desempeño del trabajo. De corroborarse, esta situación podría equipararse a algunos de los resultados obtenidos con la GATB (Hartigan y Wigdor, 1989). Más aún, los coeficientes de validez del ASVAB en comparación con diversos criterios fueron lo bastante modestos para justificar la búsqueda de nuevos predictores. La segunda fase del proyecto JPM explora modelos para los estándares de reclutamiento que aumenten la utilidad global de las decisiones de selección y clasificación, tanto en términos de costos como de niveles de desempeño.

El Proyecto de Selección y Clasificación del Ejército estadounidense (Proyecto A) incluye otro segmento importante de la investigación sobre la ASVAB y sobre nuevos predictores del desempeño de trabajos militares. El proyecto A se originó como una respuesta del ejército al mandato del proyecto JPM y puede ser el más grande y más costoso proyecto de investigación de selección de personal (Schmidt *et al.*, 1992). Además de su contribución a la formulación de una teoría del desempeño laboral (que vimos en una sección anterior), el trabajo del proyecto A ha incluido el desarrollo y la evaluación de muchos nuevos predictores que van más allá de las funciones cognoscitivas tradicionales de la ASVAB. La batería en evolución incluye pruebas psicomotoras y perceptuales aplicadas por computadora, inventarios dedicados a evaluar intereses y variables de personalidad, temperamento y antecedentes (McHenry *et al.*, 1990; N. G. Peterson *et al.*, 1990). Más aún, la gran escala y la naturaleza longitudinal del proyecto A también han hecho posible estudiar a un nivel sin precedente los cambios que ocurren con el tiempo en la validez (J. P. Campbell, 1990b).[30]

Pruebas de aptitudes especiales. Incluso antes de la aparición de las baterías de aptitudes múltiples, muchos reconocían que la cobertura que los tests de inteligencia hacían de las habilidades era limitada. Pronto empezaron a realizarse esfuerzos para llenar las principales brechas mediante tests de aptitudes especiales que abarcaban las habilidades más concretas y prácticas, como la aptitud mecánica. De igual modo, la demanda de selección de personal y de consejería estimularon el establecimiento de medidas de aptitudes espaciales, para el trabajo de oficina, musical y artística. Por otro lado, en la selección y clasificación de personal para fines industriales y militares, también se difundió el uso de pruebas de visión, audición, desempeño muscular y destreza motora.[31]

---

[29] El trabajo conceptual y metodológico que se realizó al desarrollar las pruebas basadas en el desempeño fue una valiosa contribución del proyecto JPM que seguramente demostrará su utilidad, entre otros, en el campo de la evaluación educativa.

[30] Para finales de los noventa está programada la publicación del informe final de la investigación del Proyecto A en un volumen compilado por J. P. Campbell y D. Knapp.

[31] J. C. Hogan (1992) presenta una revisión general de los procedimientos de selección de personal específicamente para trabajos de esfuerzo físico.

Respecto al concepto de *aptitudes especiales*, debe agregarse que el término se originó cuando la evaluación hacía mayor énfasis en la inteligencia general; por consecuencia, en la descripción del individuo se consideraba que aptitudes especiales, como la mecánica y la musical, entre otras, eran complementarias al "CI". Sin embargo, con la aparición del análisis factorial se fue reconociendo que la propia inteligencia comprende una serie de aptitudes relativamente independientes, como la comprensión verbal, el razonamiento numérico, la visualización espacial, etc. Más aún, ahora se incorporan a ciertas baterías de aptitudes múltiples algunas de las aptitudes especiales tradicionales, como la mecánica y la aptitud para el trabajo de oficina.

¿Cuál es entonces la función de los tests de aptitudes especiales? Primero, hay áreas como la visión, la audición, la destreza motora y el talento artístico que rara vez se incluyen en las baterías de aptitudes especiales: las situaciones que requieren la evaluación de estas áreas son demasiado especializadas para justificar la inclusión de pruebas en las baterías estándares. No obstante, los tests de aptitudes especiales también se emplean en áreas que son cubiertas en las baterías de aptitudes múltiples, como la mecánica y la del trabajo de oficina.

En algunos programas de evaluación se combinan los tests de inteligencia académica con pruebas seleccionadas de otras aptitudes relevantes. Ello obedece, entre otras cosas, a que se dispone de muchos datos normativos y de validación para algunas de las pruebas de mayor uso de aptitudes especiales. Sin duda, otra razón es la flexibilidad que este procedimiento ofrece, no sólo en la elección de las aptitudes relevantes, sino también en la plenitud con que se mide cada aptitud para propósitos específicos.

Se han desarrollado muchas pruebas para medir la rapidez, la coordinación y otras *habilidades psicomotoras*. La mayor parte de ellas se relaciona con la destreza manual, pero algunas incluyen movimientos de piernas y pies que pueden requerirse para realizar ciertos trabajos. Otras miden una combinación de aptitudes motoras y perceptuales, espaciales o mecánicas. La principal aplicación de esos instrumentos ha sido en la selección de personal industrial y militar. Las psicomotoras suelen ser pruebas con aparatos, aunque se han diseñado algunas adaptaciones de lápiz y papel para la aplicación colectiva. Un ejemplo de un instrumento publicado que requiere diversas habilidades simples de manipulación es el Test de Destreza con Piezas Pequeñas de Crawford (*Crawford Small Parts Dexterity* Test, Crawford y Crawford, 1981), que se muestra en la figura 17.2. En la primera parte, el examinado utiliza pinzas para insertar clavijas en huecos ajustados y luego coloca un pequeño aro sobre cada clavija. En la segunda, coloca pequeños tornillos en huecos roscados y luego los introduce con un destornillador. La calificación es el tiempo requerido para completar cada parte.

¿Qué puede decirse sobre la eficacia de las pruebas psicomotoras en conjunto? El punto más importante a notar al evaluarlas es el alto grado de *especificidad* de las funciones motrices. Las correlaciones y los análisis factoriales de grandes números de pruebas motoras no han logrado revelar grandes factores de grupo, como los encontrados para las funciones intelectuales (Fleishman, 1975; Fleishman y Quaintance, 1984, capítulo 12). Al considerar la validez de las pruebas psicomotoras es necesario distinguir entre las pruebas motoras complejas que se parecen al criterio particular de ejecución que pretenden predecir y las de funciones motrices simples

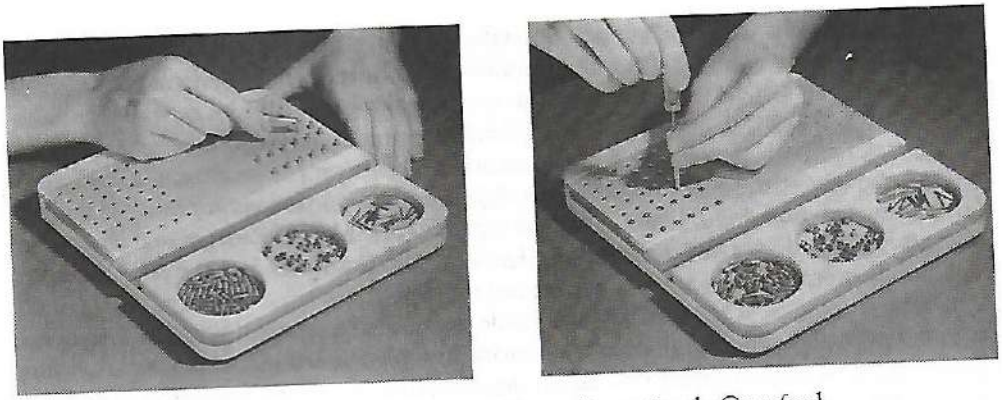


Figura 17.2. Prueba de Destreza con Partes Pequeñas de Crawford. (Cortesía de The Psychological Corporation.)

diseñadas para un uso más general. Las primeras son instrumentos hechos especialmente (que en la actualidad suelen ser computarizadas), que reproducen la combinación de aptitudes motoras requeridas por el criterio y han demostrado buena validez; por ejemplo, se ha demostrado que algunas pruebas de la fuerza aérea estadounidense mejoran la predicción de la ejecución de vuelo (véase, por ejemplo, R. H. Cox, 1989; Kantor y Carretta, 1988). Sin embargo, para casi todo propósito el uso de estas pruebas es poco práctico porque se han desarrollado muchos de esos instrumentos para igualar criterios diferentes. En lo que respecta a las pruebas motoras disponibles, las funciones que miden son muy simples y, por lo general, no es muy elevada su validez comparada con la mayor parte de los criterios, por lo que pueden funcionar mejor como parte de una batería de selección que como predictores por separado.

Las *pruebas de aptitud mecánica* cubren una diversidad de funciones. En algunos de los instrumentos de esta categoría se incluyen los factores psicomotores, ya sea porque en la ejecución del instrumento se requiere la rápida manipulación de materiales o bien porque en la prueba de lápiz y papel se incluyen subtests especiales diseñadas para medir la destreza motora. Las aptitudes perceptuales y espaciales también tienen una parte importante en muchos de estos instrumentos. Por último, el razonamiento mecánico y la información mecánica pura predominan en determinadas pruebas de aptitud mecánica.

Es importante reconocer la diversidad de funciones reunidas bajo el encabezado de aptitud mecánica, ya que cada función puede estar diferencialmente relacionada con otras variables; por ejemplo, las pruebas de información mecánica dependen mucho más de la experiencia con objetos mecánicos que las perceptuales o espaciales abstractas. De igual modo, las diferencias de género pueden ser revertidas de una a otra función, de modo que en las pruebas de destreza manual y de discriminación perceptual las mujeres suelen destacar, en las pruebas espaciales abstractas suele encontrarse una diferencia promedio pequeña pero significativa a favor de los hombres, mientras que en las pruebas de información o razonamiento mecánico éstos son marcadamente superiores (Anastasi, 1981c; Hedges y Nowell, 1995).

La aptitud espacial se incluye en todas las baterías de aptitudes múltiples. Ésta es la habilidad medida por la prueba de Relaciones Espaciales del DAT (véase el capítulo 10), y se ha encontrado que tiene un paso elevado en muchas pruebas de ejecución y sin lenguaje de inteligencia general. Una de las mejores medidas simples de esta aptitud es el Tablero de Formas de Papel de Minnesota, Revisado (*Revised Minnesota Paper Form Board Test*, RMPFBT; Likert y Quasha, 1995). Otra importante prueba de aptitud mecánica está relacionada con la información, el razonamiento o la comprensión mecánica. Aunque exige cierta familiaridad con las herramientas comunes y las relaciones mecánicas, estos instrumentos no suponen más conocimiento técnico del que puede ser adquirido en la experiencia cotidiana en una sociedad industrializada moderna. Algunas de las primeras pruebas en el campo requerían que el examinado armara objetos comunes con las piezas que le eran proporcionadas. Para propósitos generales de examinación, ahora suelen emplearse las pruebas colectivas de lápiz y papel. Un ejemplo bien conocido es el Test de Comprensión Mecánica de Bennett (*Bennett Mechanical Comprehension Test*, G. K. Bennett, 1994). Al utilizar dibujos con preguntas cortas que deben ser contestadas (como las ilustradas en la figura 17.3), la prueba hace hincapié en la comprensión de los principios mecánicos aplicados a una amplia variedad de situaciones cotidianas.

Las pruebas diseñadas para medir *aptitudes para el trabajo de oficina* se caracterizan por un énfasis común en la velocidad perceptual y la precisión. El ejemplo clásico es el

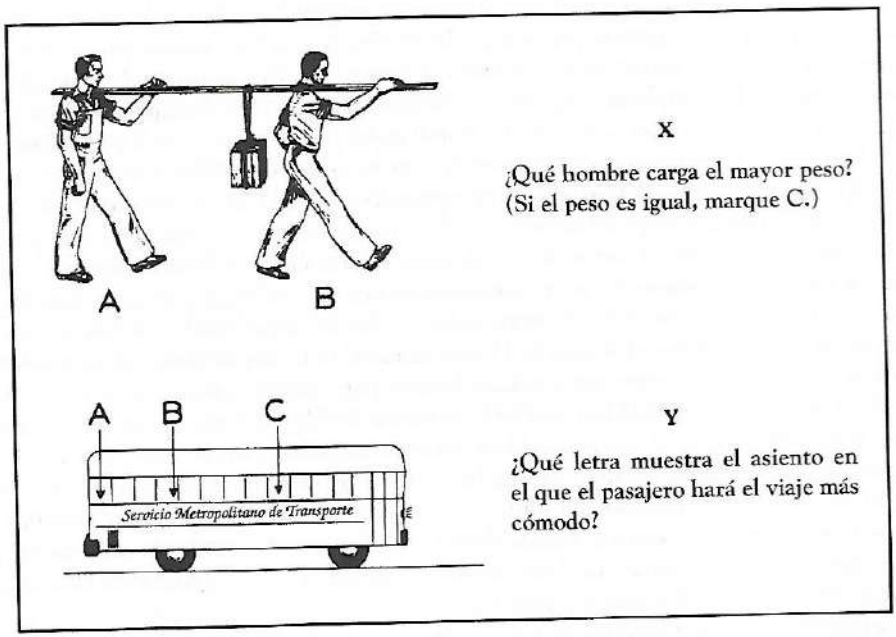


Figura 17.3. Reactivos de muestra del Test de Comprensión Mecánica de Bennett. Las respuestas se registran en una hoja de respuestas separada.

(Reproducido con autorización. Copyright © 1967–1970, 1994 por The Psychological Corporation. Todos los derechos reservados.)

Test para Empleados de Oficina de Minnesota (*Minnesota Clerical Test*, MCT, Andrew, Paterson y Longstaff, 1979), que consta de dos pruebas cronometradas por separado: comparación de números y comparación de nombres. En la primera, el examinado recibe 200 pares de números, cada uno de los cuales contiene de tres a 12 dígitos; si los dos números del par son idénticos, coloca una marca entre ellos. La tarea es similar en el segundo subtest, en el que los nombres propios sustituyen a los números.

Instrumentos relativamente homogéneos como el MCT miden sólo un aspecto del trabajo de oficina, que por su parte abarca una multiplicidad de funciones. Más aún, el número y la combinación de tareas varía enormemente con el nivel y el puesto; sin embargo, pese a semejante diversidad de actividades, los análisis de puestos del trabajo de oficina suelen indicar que se dedica una parte importante del tiempo a tareas que requieren rapidez y exactitud en la percepción de los detalles. Es indudable que, además del trabajo de oficina, muchos otros empleos requieren de rapidez y exactitud perceptual. Es obvio que inspectores, supervisores, empacadores y demasiados obreros necesitan esta habilidad, aunque muchas de estas labores las están realizando ahora dispositivos electrónicos de identificación.

Algunos tests de aptitud para el trabajo de oficina combinan la rapidez y la exactitud perceptual con otras funciones requeridas para este trabajo. Entre los instrumentos usados están las pruebas de muestras de trabajo para actividades como alfabetizar, clasificar, codificar, etc. Además, pueden incluirse algunas medidas de la habilidad verbal y numérica en lugar de un test de inteligencia general. Otros tests de aptitud para el trabajo de oficina incluyen habilidades como vocabulario e información de negocios, ortografía y manejo del lenguaje. Algunas pruebas podrían clasificarse con mayor precisión como instrumentos de aprovechamiento o muestras de trabajo, ya que miden habilidades adquiridas después de completar la capacitación especializada (por ejemplo, las pruebas de taquigrafía y mecanografía), así como las habilidades para ingresar y recuperar datos como el Test CRT de Habilidades de *Science Research Associates* (SRA, 1990).

El rápido crecimiento del uso de las computadoras en el trabajo de oficina dio lugar a la publicación de varias pruebas para *aptitudes relacionadas con el uso de las computadoras*, muchas de las cuales fueron diseñadas para la consejería o la selección de individuos que podrían ser asignados a capacitación. Entre éstas se incluyen, por ejemplo, las Pruebas de Conocimiento de las Computadoras y de la Ciencia de la Computación, así como la Batería de Aptitudes para la Programación de Computadoras (*Computer Science Test* y *Computer Programmer Aptitude Batery*; para una revisión de esos instrumentos, véase Mahurin, 1992; Marco, 1992; Schafer, 1992). Si bien en el momento en que se desarrollaron estos instrumentos (entre la década de los sesenta y el inicio de los años ochenta) representaban una aplicación oportuna de la psicometría a la evaluación de personal, el ritmo del cambio tecnológico es tan rápido que algunos de ellos se han vuelto ya obsoletos. Gradualmente se dispone de nuevos instrumentos para la instrucción y la evaluación de la competencia en el uso de programas de cómputo como dBASE, WordPerfect y Lotus 1-2-3.[32]

Es probable que algunos esfuerzos recientes por reconocer la importancia de las dimensiones cognoscitivas del funcionamiento interior y exterior del individuo (véase, por ejemplo, H. Gardner, 1983; Salovey y Mayer, 1990) estimulen el desarrollo de ins-

[32] Estas pruebas y otros instrumentos relacionados con las computadoras pueden obtenerse del SRA Product Group of McGraw-Hill/London House (véase el apéndice B).

trumentos objetivos estandarizados para la evaluación de los *aspectos sociales y emocionales de la inteligencia* en el contexto laboral. Hasta ahora, estos instrumentos prácticamente no existen, en la medida en que lo adecuado de tales funcionamientos en los ambientes laborales tradicionalmente han sido medidos mediante tests de personalidad, entrevistas o técnicas del centro de evaluación. Uno de los primeros instrumentos en esta nueva categoría es el Test KSA para el Trabajo en Equipo (*Teamwork–KSA Tests*), una prueba de lápiz y papel recién publicada y destinada a predecir la eficacia en el trabajo de equipo. El instrumento, que se fundamentó en una revisión de la bibliografía sobre los grupos de trabajo en las organizaciones, pretende medir el conocimiento, las habilidades y capacidades de los examinados en los campos personal y de autocontrol. Los reactivos de opción múltiple presentan preguntas hipotéticas que atañen a la solución de conflictos, la comunicación y la resolución de problemas mediante la colaboración, así como el establecimiento de metas, planeación y otras habilidades de automanejo (M. J. Stevens y Campion, 1994). Es de esperarse que prosiga el trabajo y la experimentación con éste y otros instrumentos similares.

Evaluación de la personalidad en el lugar de trabajo

A mediados de los ochenta, Bernardin y Bownas (1985) advirtieron que en los medios industriales florecía el uso de técnicas de evaluación de la personalidad (incluido el uso de métodos no científicos como el análisis de la escritura), mientras que la comunidad académica había descuidado el tema durante casi dos décadas.[33] Desde entonces, ha tenido lugar un considerable resurgimiento del trabajo en el área, influido por los desarrollos teóricos y metodológicos. Desde un punto de vista metodológico, la aplicación del meta-análisis y de las técnicas de modelamiento causal (capítulo 5) han estimulado la investigación de los rasgos no cognoscitivos que pueden afectar el desempeño en el trabajo. Los meta-análisis se han empleado principalmente para investigar la validez y utilidad de los constructos de personalidad entre diversos ambientes. El análisis de trayectorias y las técnicas de modelamiento de ecuaciones estructurales (capítulo 5) se emplean para estudiar las correlaciones entre los predictores y para explorar los atributos y las condiciones que dan lugar a diferentes grados de desempeño en el trabajo. También se está estudiando el grado al que algunas variables cruciales, como el nivel de habilidad y la autonomía en el trabajo, moderan la relación entre personalidad y desempeño. En general, las metas de esas investigaciones superan la mera predicción y se dirigen a la comprensión de los constructos y procesos que determinan las grandes variaciones en el desempeño laboral (Barrick y Mount, 1991, 1993; Hough, Eaton, Dunnette, Kamp y McCloy, 1990; Matthews, Jones y Chamberlain, 1992; McHenry *et al.*, 1990; Schmidt y Hunter, 1992; Tett, Jackson y Rothstein, 1991).

La investigación de metodología más compleja que ha tenido lugar en las dos últimas décadas ha proporcionado apoyo considerable al uso de tests de personalidad cuidadosamente construidos en la toma de decisiones relativas al empleo en diversos entornos. Un aspecto importante que requiere de mayor investigación es la determi-

[33] Véase Anastasi (1985e) para una revisión histórica de la evaluación de la personalidad en la industria y para un análisis de diversos problemas metodológicos y prácticos básicos relacionados con ella. Kanfer *et al.* (1995) hicieron una revisión del progreso en la aplicación de los conceptos de las teorías de la personalidad y de la inteligencia al campo de la psicología industrial y de las organizaciones.

nación de las dimensiones de personalidad más relevantes para el desempeño de ciertos trabajos o familias de trabajos; por ejemplo, mientras que algunos empleos pueden requerir una elevada sociabilidad, en otros puede ser el caso contrario. Incluso una característica como la escrupulosidad, que a primera vista parece deseable para cualquier empleo, al analizarse más de cerca puede revelar que, al menos en ciertos aspectos, incluye atributos contraproducentes para algunos empeños, como las actividades creativas. Como resultado de estas consideraciones, se empieza a prestar atención al establecimiento de los requisitos temperamentales e interpersonales de los trabajos (R. Hogan, Hogan y Roberts, 1996; Landy *et al.*, 1994).

Gran parte de la actividad reciente de investigación se ha concentrado en explorar la utilidad del Modelo de los Cinco Factores de la estructura de la personalidad, en sus distintas formas, para la predicción de criterios de desempeño en el trabajo (véase el capítulo 13 y R. Hogan, 1991). La popularidad del modelo en la investigación sobre la selección de personal no es accidental; después de todo, los cinco (± 2) factores fueron extraídos de un vasto arreglo de datos descriptivos de la personalidad y representan un modelo para la descripción de las personalidades normales respecto al cual existe considerable acuerdo. Cada uno de los dominios incluidos en el modelo es aplicable a cualquier conducta cotidiana, incluido el desempeño de las tareas del trabajo; por ejemplo, la dimensión de la Estabilidad Emocional resulta esencial en los puestos que exigen que se tomen decisiones rápidas en condiciones de presión, como los mandatos legales, la conducción de aviones y camiones y el trabajo en la sala de emergencias médicas. Por otro lado, la Agradabilidad es una cualidad indispensable para cualquier trabajo que incluye un intenso contacto con los demás. Con todo, no resulta sorprendente que el factor de Escrupulosidad sea el que ha aparecido en varias investigaciones meta-analíticas como el predictor de personalidad más universal e importante del desempeño en el trabajo (Barrick y Mount, 1991, 1993; Schmidt y Hunter, 1992). Aun así, el acuerdo está lejos de ser universal con respecto a la definición del constructo de Escrupulosidad o a la cuestión de su primacía (véase, por ejemplo, Loevinger, 1994; Tett *et al.*, 1991).

Pruebas de integridad. La aplicación de pruebas de integridad u honestidad para tomar decisiones sobre el empleo adquirió especial importancia luego de la aprobación de la Ley de Protección a Empleados contra el Polígrafo, de 1988 (P. L. 100–347), que prohibía el uso de polígrafos en la selección de personal, salvo en situaciones especiales como en las dependencias gubernamentales.[34] Como consecuencia, proliferaron las mediciones de lápiz y papel sobre la honradez, conocidas como "pruebas de integridad". Estos instrumentos, que por lo general indagan sobre las actitudes y los antecedentes del examinado en robos y otras conductas ilícitas, rápidamente fueron sometidas a escrutinio dentro y fuera de la psicología.[35] En una de las más extensas investiga-

[34] Una de las primeras revisiones críticas del tema de la detección de mentiras escrito por un psicólogo es el libro de Lykken (1981). Otras revisiones más breves y recientes pueden encontrarse en DePaulo (1994), Honts (1994), Kircher y Raskin (1992), Lykken (1992) y Saxe (1994).

[35] Véase, por ejemplo, el informe de una fuerza de tarea que formó la Asociación Psicológica Estadounidense para indagar sobre el tema (L. R. Goldberg, Grenier, Guion, Sechrest y Wing, 1991). Puede encontrarse información adicional en el amplio volumen dedicado por K. R. Murphy (1993) al tema de la honestidad en el lugar de trabajo. Para una revisión breve pero informativa de la evaluación de la integridad en la selección de personal, véase Cámara y Schneider (1994), Sackett (1994) y en la sección de comentarios del número correspondiente a junio de 1995 de *American Psychologist*.

ciones sobre las pruebas de integridad, Ones, Viswesvaran y Schmidt (1993) realizaron un meta-análisis basado en 665 coeficientes de validez. Su estudio estimó que la validez operacional promedio de las pruebas de integridad para predecir las valoraciones de los supervisores del desempeño laboral es de .41, e indicaron que los instrumentos también pueden ser valiosos en la predicción de la conducta desordenada en el trabajo.

A pesar del apoyo sustancial al uso de las pruebas de integridad, dentro de ciertos parámetros, su aplicación en las decisiones de selección de empleados continúa siendo problemática. La preocupación actual se concentra en la posible susceptibilidad del instrumento a la preparación y el falseamiento y en su relativa ineficacia para predecir formas específicas de conducta inadecuada en el trabajo, como los hurtos (véase, por ejemplo, Alliger, Lilienfeld y Mitchell, 1996; Cámara y Schneider, 1995).

Liderazgo. La selección de líderes eficaces plantea otro desafío importante en el campo de las decisiones relacionadas con el personal. El liderazgo es una de las cualidades más demandadas en el campo de trabajo porque abarca la habilidad de persuadir a los demás para trabajar en una causa común. El liderazgo eficaz puede tener consecuencias poderosas para mejorar el funcionamiento de la organización, mientras que el liderazgo ineficaz puede resultar desastroso. Así, no resulta sorprendente que en la selección de los ejecutivos de alto nivel se apliquen muchas clases de pruebas cognoscitivas y tests de personalidad, lo mismo que las técnicas del centro de evaluación, entrevistas, simulaciones y cuestionarios para la obtención de datos biográficos. A menudo, el proceso de evaluación de la selección de estos ejecutivos incluye diversos métodos y se extiende por varias horas o días. Los procedimientos son muy similares a los de la evaluación clínica y por lo general son muy costosos.

Los recursos dedicados a la selección cuidadosa de los individuos que pasan de la fuerza de trabajo regular a ocupar posiciones de liderazgo parecen ser relativamente magros cuando se comparan con el impacto potencial de las decisiones. Es posible que, como consecuencia de esta situación, la incompetencia gerencial parezca ser demasiado alta (R. Hogan, Curphy y Hogan, 1994). Aunque el interés por el estudio de la conducta de liderazgo tiene una larga historia en la psicología aplicada y ha tenido como resultado una extensa bibliografía sobre el tema, queda mucho por hacer en términos de identificar y aclarar los correlatos empíricos del liderazgo eficaz. Para mayor información sobre las mediciones de liderazgo y sobre los avances teóricos y metodológicos en el estudio del liderazgo en organizaciones, véase Bass (1990), Clark y Clark (1990), y Yukl y Van Fleet (1992).

Instrumentos. Con excepción de los procedimientos intensivos de evaluación empleados en la selección de ejecutivos de alto nivel y de cierto personal especializado (como agentes de inteligencia, astronautas y tareas afines), a la gran mayoría de los solicitantes de empleo que presentan pruebas de personalidad se les aplica un inventario autodescriptivo del tipo que estudiamos en el capítulo 13. Los inventarios empleados para la detección de psicopatologías, como el MMPI, todavía se emplean con propósitos de proyección en ciertas ocupaciones especiales, pero la preocupación por la invasión de la intimidad y el mal uso de los resultados ha conducido a la utilización predominante de tests diseñados para evaluar la personalidad en individuos normales (véase el capítulo 18). De hecho, varios de los antiguos inventarios multidimensionales autodescriptivos que originalmente se destinaban a personas normales —por ejemplo, el CPI—

han sido revisados recientemente para hacerlos adecuados para emplearse en primer lugar en los medios laborales. Al mismo tiempo, se diseñaron algunas medidas nuevas, como el Inventario de Personalidad de Hogan (*Hogan Personality Inventory*, HPI-R); Hogan & Hogan, 1992), en primer lugar para medios ocupacionales. Se preparan otras pruebas para una gama más restringida de aplicaciones, por ejemplo el Inventario de Servicios al Cliente PDI y el Inventario de Empleo PDI, que es un instrumento diseñado para identificar a los solicitantes que serán empleados confiables y estables (véase Paajanen, Hansen y McLellan, 1993).

Comentarios finales. Los adelantos tecnológicos están transformando la naturaleza del trabajo de formas que superan por mucho los cambios producidos por la Revolución Industrial. El rápido ritmo del cambio supone un enorme desafío para los psicólogos de la personalidad, aunque los avances metodológicos que las nuevas tecnologías han hecho posibles ofrecen grandes oportunidades.[36] Justo cuando se ha hecho factible realizar la investigación a gran escala y compleja que permite a los estudiosos obtener algunas respuestas definitivas a preguntas que habían quedado pendientes durante mucho tiempo, las propias preguntas han comenzado a cambiar. Mientras tanto, el impacto de la tecnología computarizada sobre la evaluación del personal, aunque visible, apenas inicia (Bartram, 1993; Burke, 1993; Schoenfeldt y Mendoza, 1991). Algunas de las aplicaciones más interesantes de la tecnología a la evaluación (como las pruebas interactivas por multimedios) aún se encuentran en las etapas preliminares de desarrollo (Desmarais, Masi, Olson, Barbera y Dyer, 1994; Drasgow *et al.*, 1996) y aunque pueden anticiparse muchas otras, todavía no se han concretado. Pero, en conjunto, parece haber mucho espacio para la aplicación meditada de los principios y métodos psicológicos para mejorar la utilización de los recursos humanos (Bray *et al.*, 1991; Gatewood y Feild, 1993).

USO DE LAS PRUEBAS EN LA PSICOLOGÍA CLÍNICA Y LA CONSEJERÍA

Los psicólogos clínicos y los consejeros emplean una amplia variedad de pruebas, entre las que se incluyen la mayor parte de las que hemos visto. Periódicamente se realizan estudios sobre el uso de las pruebas por muestras de psicólogos clínicos y consejeros que difieren en su orientación teórica y sus lugares de trabajo (Archer, Maruish, Imhof y Piotrowski, 1991; Lubin, Larsen y Matarazzo, 1984; Piotrowski y Keller, 1992; Watkins *et al.*, 1994; Watkins, Campbell, Nieberding y Hallmark, 1995). Entre las de mayor uso están los tests de inteligencia individual, como las escalas de Wechsler y baterías similares de aptitudes múltiples como el DAT (véase el capítulo 10). Muchás de los tests de personalidad que tratamos en los capítulos 13 a 16 destacan en el repertorio de clínicos y consejeros (Butcher y Rouse, 1996); y ciertas pruebas de diagnóstico educativo son apropiadas en los casos de problemas de aprendizaje y otras dificultades escolares. Además, los clínicos y consejeros utilizan muchos cuestionarios breves y escalas de valoración para una evaluación rápida de la multiplicidad de problemas enfrentados en la práctica. Muchas de estas medidas se describen en la obra de J. Fischer y Corcoran (1994).

[36] El lector puede encontrar en Landy, Shankster-Cawley y Kohler-Moran (1995) una revisión de las transformaciones que suceden en el lugar de trabajo y las consecuencias que pueden tener para el trabajo del personal.

Esta sección se concentra principalmente en las áreas que requieren la aplicación de múltiples pruebas y otros instrumentos para la evaluación de individuos, en contraste con el uso aislado de un único instrumento. Las evaluaciones psicológicas de esta clase son conducidas por psicólogos de diversos campos, aunque muchos se han formado en la psicología clínica y la consejería así como en sus subespecialidades. Por tradición, los psicólogos clínicos realizan evaluaciones con propósitos de diagnóstico, pronóstico y toma de decisiones terapéuticas en medios de salud mental (Butcher, 1995; Hersen, Kazdin y Bellack, 1991; Hurt, Reznikoff y Clarkin, 1991; Maruish, 1994), aunque al principio los psicólogos consejeros se especializaban en exclusiva en la evaluación y la orientación vocacional (S. D. Brown y Lent, 1992; Drummond, 1996; Gelso y Fretz, 1992). Otros campos importantes en los que se practica la evaluación psicológica es la psicología forense o legal,[37] que deriva su nombre de los ambientes en los que trabajan los profesionales (Shapiro, 1991; Vance, 1993; Weiner y Hess, 1987). Otras especialidades que utilizan las técnicas de evaluación psicológica las definen las clases de problemas que estudian, como la psicología de la salud (N. Adler y Matthews, 1994; Streiner y Norman, 1995; S. E. Taylor, 1990) y la neuropsicología, un campo interesado en las relaciones entre cerebro y conducta (Benton, 1994; Maruish y Moses, 1997). Otros son identificados por los pacientes o clientes a los que atienden, por ejemplo niños, parejas o familias (Conoley y Werth, 1995; Fruzzetti y Jacobson, 1992; Kamphaus y Frick, 1996). Muchos trabajan en áreas todavía más especializadas, como la neuropsicología pediátrica (Batchelor y Dean, 1996) o la rehabilitación médica (Cushman y Scherer, 1995). Todos estos campos y subespecialidades evolucionan constantemente; además, los límites entre algunas especialidades han empezado a ser menos claros y, en algunos casos, casi han desaparecido.

Un caso notable de este traslape en la aproximación creciente entre la psicología clínica y la consejería (véase, por ejemplo, Anastasi, 1979, 1990a; May, 1990).[38] Cada vez es más frecuente encontrar a los psicólogos consejeros trabajando en la práctica privada, las clínicas y los hospitales; al mismo tiempo, algunos de ellos aplican ahora sus habilidades en la evaluación profesional y la selección de personal (Lowman, 1989, 1991, 1993). En un signo visible de los cambios que tienen lugar, los *Estándares de Examinación* recientemente propuestos abordan el uso de las pruebas para el diagnóstico y evaluación individual en los medios clínico, de consejería y escolar en un solo capítulo titulado *Psychological Testing and Assessment* ("Examinación y Evaluación Psicológica"), mientras que la versión de 1985 aborda por separado la evaluación clínica y el uso de las pruebas en la consejería (véase el capítulo 1).

Evaluación psicológica. ¿Qué es entonces lo que vincula a la práctica de la evaluación psicológica en diferentes especialidades, escenarios, áreas problemáticas y

[37] La psicología forense o la legal es la práctica de la psicología en los tribunales en relación con asuntos legales o con individuos juzgados o encarcelados.

[38] Debido a esa circunstancia, así como por conveniencia, en esta sección se utilizará el término "clínico" para denotar a cualquier profesional que practica la evaluación psicológica. No obstante, resulta útil recordar que los antecedentes educativos de estos profesionales pueden corresponder a cualquiera de varias especialidades y, por ende, es posible que difieran en su orientación y áreas de experiencia. Véase P. S. Wise (1989) para obtener mayor información sobre el uso que los psicólogos clínicos, consejeros, industriales/organizacionales y escolares hacen de las pruebas en sus entornos tradicionales.

poblaciones? Una de sus características más importantes es el interés en el estudio exhaustivo de uno o más individuos (por ejemplo, una pareja o una familia) mediante numerosas fuentes de datos. Al establecer y mantener el *rapport*, el clínico puede obtener del cliente los hechos pertinentes de su historia a los que no es fácil acceder por otros medios. Estos datos ofrecen una base particularmente adecuada para comprender al individuo y predecir su conducta. Más aún, el clínico contribuye al proceso de encontrar los hechos al fungir como estímulo en una situación de trato personal. A este respecto, la entrevista clínica funciona como prueba situacional o simulación que ofrece una muestra de la conducta del cliente en condiciones más o menos controladas.

La información derivada de la observación, las entrevistas y el historial del caso se combina con las calificaciones obtenidas en la prueba para ofrecer una imagen integrada del individuo (Beutler y Berren, 1995). El clínico dispone así de ciertas salvaguardas contra la generalización excesiva de las calificaciones de la prueba, lo que probablemente explique al menos en parte que continúe el uso de algunas pruebas cuyas propiedades psicométricas son débiles o no demostradas. Su retención puede justificarse en la medida en que dichos instrumentos sirven ante todo para sugerir a los clínicos experimentados el rumbo a seguir; por supuesto, existe el riesgo de que un clínico relativamente inexperto y demasiado celoso, al no considerar las limitaciones de un instrumento deposite más confianza en sus resultados de lo que está justificado.

Otro atributo que define a la evaluación psicológica es su objetivo, que por lo general consiste en ayudar a tomar decisiones relativas al diagnóstico diferencial, la selección de carrera, los tratamientos recomendados, la planeación educativa, las determinaciones sobre la custodia de los hijos, la culpabilidad y muchos otros asuntos de significado práctico para una o más personas. La toma de decisiones procede por medio de la obtención, el análisis, la integración y el informe cuidadoso de los datos conductuales relevantes. En el centro del proceso de la evaluación psicológica se encuentra una secuencia continua de generación y comprobación de hipótesis sobre el caso individual. Cada parte de información, se trate de un acontecimiento registrado en el historial del caso, el comentario de un cliente o la calificación de una prueba, propone una hipótesis acerca del individuo que será confirmada o refutada al obtener información sobre otros hechos. También a este respecto debe tenerse presente que cualquier fuente aislada de datos, sin importar lo confiable que parezca, en ocasiones puede proporcionar información inexacta.[39]

Fundamentalmente, todas las actividades relacionadas con la evaluación psicológica (de la clara elucidación de su propósito principal a la comunicación de los resultados) implican el *juicio* profesional basado en el conocimiento sobre los problemas específicos y las poblaciones a la mano. Más aún, la aplicación de las pruebas individuales y otras herramientas de evaluación requieren de habilidades especializadas en su uso y de la cuidadosa consideración de sus características a la luz del propósito y el contexto de la evaluación (Drummond, 1996; C. T. Fischer, 1985; Groth-Marnat, 1990; G. Goldstein y Hersen, 1990; Hood y Johnson, 1997; Maloney y Ward, 1976; Tallent, 1992; Walsh y Betz, 1995). A este respecto, el método clínico ha sido comparado con la aplicación de pruebas objetivas estandarizadas y con el uso de procedimientos estadísti-

[39] En un número especial de *Psychological Assessment* (septiembre 1995, vol 7, núm. 3) puede encontrarse una muestra representativa de los temas metodológicos actuales en la investigación de la evaluación psicológica.

cos o actuariales para combinar datos mediante ecuaciones de regresión, cortes múltiples y otras "fórmulas automáticas" (capítulo 6; véase también Dawes, Faust y Meehl, 1993; L. R. Goldberg, 1991; Kleinmuntz, 1990; Wedding y Faust, 1989).

El punto de vista ecológico, que insiste en la necesidad de tener en cuenta el contexto de la vida del individuo, ha tenido una influencia considerable en la psicología del desarrollo y otros campos afines (véase, por ejemplo, Moen, Elder y Lüscher, 1995). De igual modo, la mayor conciencia de la función que cumple la cultura en toda la conducta (incluidos los problemas que llevan a la gente a buscar ayuda de los profesionales de la salud mental) ha estimulado un gran interés en la información y las directrices para la práctica culturalmente competente de la evaluación (véase, por ejemplo, American Psychiatric Association, 1994, pp. xxiv y 843–849; Dana, 1993, 1996; Prediger, 1993; Suzuki *et al.*, 1996).

Los tests de inteligencia en el contexto de la evaluación individual. Pruebas como las escalas de Wechsler y la de Stanford-Binet (véase el capítulo 8) son en esencia instrumentos clínicos de aplicación individual. Cuando un clínico alerta y capacitado está en contacto activo con un examinado durante —aproximadamente— la hora que se requiere para aplicar la prueba, es difícil que no obtenga más conocimiento de la persona que el transmitido por el CI o cualquier otra calificación aislada. Así ocurre incluso si el instrumento es aplicado por un técnico, toda vez que se conserva el registro completo de las respuestas del examinado.

Además de utilizar los tests de inteligencia para evaluar el nivel general del funcionamiento intelectual del individuo, los clínicos también acostumbran explorar el patrón, o perfil, de las calificaciones en búsqueda de las fortalezas o debilidades significativas. El análisis del perfil proporciona datos que pueden ser de ayuda en el diagnóstico del daño cerebral y de varias formas de psicopatología que afectan diferencialmente el funcionamiento intelectual. Las escalas de Wechsler son en particular adecuadas para este análisis, ya que las puntuaciones de todos los subtests se expresan en calificaciones estándares comparables directamente. Desde el inicio, Wechsler describió una serie de usos diagnósticos de sus escalas, y, a partir de entonces, varios clínicos han recomendado técnicas adicionales y el análisis de perfiles se ha aplicado también con otros instrumentos (Delaney y Hopkins, 1987; Elliott, 1990b; Kaufman, 1990, 1994; Matarazzo, 1972; Sattler, 1988, 1992). La mayor parte de las técnicas de análisis de perfiles emplea variantes de tres procedimientos principales. El primero incluye la evaluación de la cantidad de *dispersión* o el grado de variación entre las diversas calificaciones del individuo; comprende las diferencias en el CI Verbal y de Ejecución, la dispersión total de los subtest y la comparación de las puntuaciones escaladas de los subtests individuales con la media de varios agrupamientos de puntuaciones de subtests, como los verbales o los de velocidad. El segundo procedimiento consiste en analizar los rasgos destacados del perfil del individuo en función de la *tasa base de datos* sobre la frecuencia o infrecuencia de los mismos en el grupo normativo. El tercer método se basa en los *patrones de calificación* asociados con ciertos síndromes clínicos, como la enfermedad de Alzheimer, los problemas de aprendizaje o los estados de ansiedad. Wechsler y otros investigadores describieron patrones de calificaciones elevadas y bajas en los subtests así como de combinaciones de subtests que caracterizan a éstos y otros trastornos (véase, por ejemplo, Kaufman, 1990; Matarazzo, 1972).

Varias décadas de investigación sobre las diversas formas de análisis da patrones con las escalas de Wechsler han ofrecido poco apoyo estadístico a sus valor diagnóstico.[40] De hecho, los críticos de esta aproximación han atacado casi todos sus aspectos en un punto u otro (F. C. Goldstein y Levin, 1985; Kavale y Forness, 1984; Macmann y Barnett, 1994a; McDermott, Fantuzzo y Glutting, 1990). No obstante, a juzgar por la continua popularidad de las escalas de Wechsler en la práctica clínica y de la enorme bibliografía destinada a sistematizar, facilitar y mejorar la utilización del análisis de patrón de calificaciones, es claro que sigue siendo la aproximación preferida para interpretar los datos de los tests de inteligencia.

A un nivel puramente cualitativo, cualquier irregularidad en la ejecución también puede sugerir nuevas rutas por explorar, y pueden surgir indicios importantes tanto de la forma como del contenido de las respuestas a la prueba; por ejemplo, la extravagancia, la sobreelaboración o una autorreferencia excesiva pueden indicar trastornos de personalidad. Un análisis cualitativo tanto de los errores como de las respuestas correctas puede ofrecer señales útiles sobre las aproximaciones a la solución de problemas, el desarrollo conceptual o los estilos cognoscitivos; asimismo, en el contenido atípico de las respuestas se encuentra una fuente adicional de indicios. Durante la aplicación de un test de inteligencia individual puede encontrarse una rica fuente de datos cualitativos en la conducta general del examinado durante la situación de prueba. Algunos ejemplos incluyen la actividad motora, el habla, las respuestas emocionales, las actitudes hacia el examinador, así como la forma de aproximarse a los materiales y al ambiente de la prueba. Como regla, y debido a su naturaleza idiosincrática, esas medidas cualitativas no pueden validarse por métodos cuantitativos adaptados a la medición de tendencias de grupo. No obstante, al reconocer la importante función que pueden cumplir, empiezan a aparecer algunos instrumentos diseñados para sistematizar y cuantificar el registro y la interpretación de algunas conductas durante la sesión de prueba (véase, por ejemplo, la Guía para la evaluación de la conducta en la sesión de prueba para el WISC-III y el WIAT de Glutting y Oakland, 1992).

Alan S. Kaufman ofrece un claro ejemplo del complejo uso clínico de los tests de inteligencia que combina datos psicométricos con las observaciones cualitativas. En su libro sobre la evaluación "inteligente" y la inteligencia (1979, 1990, 1994), demuestra en detalle la forma en que el clínico puede integrar la información estadística sobre las puntuaciones en las pruebas con el conocimiento sobre el desarrollo humano, la teoría de la personalidad y otras áreas de investigación psicológica. Kaufman recalca la importancia de considerar tanto las habilidades como las condiciones extrañas que pueden influir en la ejecución en los subtests como la necesidad de contar con información complementaria (obtenida de otros instrumentos, del historial del caso y de la observación clínica de la conducta durante la prueba) contra el que puedan interpretarse los patrones de puntuaciones. Las calificaciones obtenidas en la prueba, junto con otras fuentes de datos, conducen a la formulación de hipótesis sobre el individuo que pueden ser sometidas a prueba conforme se obtiene mayor información para redondear la imagen. La característica más importante del método de Kaufman es que requiere interpretaciones individualizadas de la ejecución en la prueba en lugar de

[40] El lector puede encontrar en Anastasi (1985a) una revisión de algunos de los requisitos metodológicos que deben considerarse al evaluar esta investigación.

cualquier análisis de patrones, ya que el mismo patrón de puntuaciones puede conducir a interpretaciones muy diferentes para distintos individuos.

Sin duda, el planteamiento de Kaufman representa una contribución importante al uso clínico de los tests de inteligencia (véase, por ejemplo, Roecker, 1995). Incluso sus críticos reconocen que se ha convertido en el método a elegir para la enseñanza de la evaluación de la inteligencia y que ha orientado la creación de buena parte de los programas de cómputo disponibles para apoyar en la interpretación de los tests de inteligencia (McDermott, Fantuzzo, Glutting, Watkins y Baggaley, 1992).[41] También las directrices ofrecidas por Sattler (1988, 1992) son excelentes ejemplos del uso clínico y psicométrico combinado de los tests de inteligencia individual. Empero, debe reconocerse que la implantación de estas aproximaciones requiere un clínico bien capacitado, informado en diversos campos de la psicología y que, por añadidura, no tenga presiones de tiempo. Más aún, aunque en cierta forma facilitan la aplicación de estos métodos, el acceso a los programas de interpretación computarizados puede llevar al evaluador apresurado o menos informado a interpretar en exceso los datos proporcionados por los programas.

Mientras tanto, se están desarrollando nuevos procedimientos para la utilización de los perfiles de los tests de inteligencia, algunos de los cuales combinan elementos de las posturas psicométrica y clínica. Una innovación interesante que está siendo explorada es la aproximación del "perfil central". McDermott, Glutting y sus colaboradores (Glutting, McDermott, Prifitera y McGrath, 1994, 1995; McDermott, Glutting, Jones y Noonan, 1989), así como Donders (1996), han aplicado numerosos análisis —de varios tipos— a la estandarización de los datos de las escalas de Wechsler. La meta de estos análisis es generar perfiles básicos que puedan ayudar a clasificar los resultados de la prueba y a comprobar las hipótesis sobre su importancia clínica. Hay una línea de investigación complementaria que utiliza el escalamiento multidimensional para identificar los perfiles prototípicos de habilidad en la población (Davison, Gasser y Ding, 1996). Esta investigación pretende cuantificar el grado de correspondencia entre el perfil individual observado y los perfiles prototípicos surgidos de determinada batería de pruebas. Aunque tales investigaciones son interesantes, aún son exploratorias y no tienen todavía utilidad clínica (véase, por ejemplo, Ryan y Bohac, 1994).

Evaluación neuropsicológica

Problemas metodológicos en el diagnóstico del daño cerebral. El conocimiento de los efectos conductuales del daño cerebral se remonta a los escritos de Kurt Goldstein y sus colaboradores a principios de los años veinte (Goldstein y Scheerer, 1941). Después de la observación prolongada de soldados que habían recibido lesiones cerebrales durante la Primera Guerra Mundial, Goldstein formuló su descripción clásica del deterioro intelectual asociado con la lesión cerebral. Entre los principales

[41] Para obtener mayor información sobre varias críticas del método de Kaufman, así como para las réplicas, véase Kaufman (1994, capítulo 1). Un problema con varias de las revisiones negativas de la aproximación de Kaufman es que parecen suponer que los clínicos lo utilizarán para tomar decisiones basados únicamente en la magnitud y las diferencias de las calificaciones. Aunque es cierto que la aplicación mecánica de las técnicas de análisis de perfiles pueden llevar a conclusiones engañosas, esta suposición es del todo contraria a las recomendaciones de Kaufman, lo mismo que a los principios de la práctica adecuada de evaluación (Moreland *et al.*, 1995).

síntomas se encuentra una disminución para el pensamiento abstracto y la tendencia a responder a los estímulos externos que pueden interrumpir la percepción normal.

La difundida preocupación por el daño cerebral en los niños surgió a finales de los treinta y en los cuarenta, gracias a la investigación de Alfred Strauss y sus colaboradores (Strauss y Lehtinen, 1947; H. Werner y Strauss, 1941, 1943), quienes identificaron a un grupo de niños con retardo mental cuyas historias de los casos mostraban evidencias de daño cerebral debido a traumas o infecciones ocurridas antes, durante o poco después del nacimiento. La descripción conductual de esos niños representó una extensión y elaboración del síndrome adulto presentado por Goldstein. Delineaba un patrón distintivo de trastornos intelectuales y emocionales que ha sido aceptado como característico "del niño" con daño cerebral. En este patrón se incluyen trastornos perceptuales y conceptuales específicos combinados con una habilidad verbal relativamente elevada, actividad excesiva, distracción y agresividad. Durante muchos años, la investigación y la práctica con niños con daño cerebral fueron dominadas por el concepto unidimensional de "organicidad". La aproximación llevó a una búsqueda de pruebas diagnósticas de implicaciones orgánicas y a la tentativa por desarrollar programas educativos o de remedio adecuados para los niños con daño cerebral en su conjunto.

Desde la década de los cincuenta, los psicólogos reconocen cada vez con más frecuencia que el daño cerebral puede conducir a una amplia variedad de patrones conductuales, reconocimiento que ha estimulado el desarrollo de la neuropsicología clínica, un campo que pretende aplicar lo que se conoce sobre las relaciones entre cerebro y conducta en el diagnóstico y rehabilitación de los individuos con daño cerebral. Ningún síntoma o conjunto de síntomas necesitan ser comunes a todos los casos de lesión cerebral; de hecho, ésta puede producir un patrón opuesto de conducta en dos individuos, descubrimientos que son consistentes con la amplia diversidad de la propia patología orgánica subyacente. Un avance significativo en el análisis de las relaciones entre cerebro y conducta fue hecho por la investigación de Ralph Reitan y sus colaboradores en el Centro Médico de la Universidad de Indiana (véase Matarazzo, 1972, capítulo 13; Reitan, 1955, 1966), en la que se demostró que las lesiones del hemisferio izquierdo tienden a asociarse con un CI Verbal menor al CI de Ejecución en las escalas de Wechsler (V < E), mientras que el patrón opuesto (V > E) predomina en los grupos con lesiones en el hemisferio derecho y con daño difuso.

La investigación neuropsicológica sigue tratando de averiguar las interacciones complejas de otras variables con los efectos conductuales de la patología cerebral (véase, por ejemplo, Kolb y Whishaw, 1990). Hay evidencias considerables de que la *edad* afecta los síntomas conductuales que resultan del daño cerebral; éstos dependen también de la cantidad del aprendizaje y desarrollo intelectual ocurridos antes de la lesión; por ejemplo, la investigación con niños preescolares indica que a este nivel de edad las lesiones cerebrales tienden a afectar el funcionamiento intelectual más masivamente que en niveles posteriores.

Se ha descubierto que también la *cronicidad* afecta la ejecución en la prueba e interactúa con los efectos de la edad. Los datos disponibles indican que el tiempo transcurrido desde la lesión puede relacionarse no sólo con los cambios fisiológicos progresivos, sino también con el grado de recuperación conductual mediante el aprendizaje o los reajustes compensatorios. Por último, debe advertirse que en algunos casos el deterioro intelectual puede ser un *resultado indirecto* del daño cerebral. A lo largo del

desarrollo del individuo, hay una interacción entre los factores orgánicos y la experiencia; por ejemplo, algunas de las dificultades conductuales incluidas en la imagen clásica de los niños con daño cerebral pueden ser un efecto indirecto de las frustraciones y problemas con las personas que sufre el niño con una deficiencia intelectual de causas orgánicas. Que las dificultades conductuales persistan o no puede depender de las actitudes y el grado de comprensión que muestren los padres, maestros y otras personas significativas en el entorno del niño.

Es evidente que el daño cerebral cubre una amplia variedad de trastornos orgánicos, con la correspondiente diversidad de manifestaciones conductuales. Puede esperarse entonces que el desempeño en la prueba de las personas con daño cerebral varíe de acuerdo con la fuente, el grado y la localización de la lesión cerebral, la edad a la que ocurrió y a la que se evalúa la conducta del individuo, la duración de la condición patológica y la extensión, así como los tipos de intervención que ha recibido. En consecuencia, resulta poco realista esperar homogeneidad conductual entre los afectados por patologías cerebrales.

Desde otro punto de vista, el mismo trastorno intelectual o conductual y el mismo signo diagnóstico de la ejecución en la prueba pueden ser resultado de etiologías orgánicas, emocionales o mixtas. Un ejemplo notable es el olvido persistente. Diversas clases de problemas de memoria pueden ser síntoma de uno de muchos tipos de demencia, con causas orgánicas conocidas, o de trastornos depresivos de origen emocional; además, el inicio de un trastorno orgánico de memoria con frecuencia está acompañado de depresión, lo que presenta un cuadro mixto. En la neuropsicología clínica, una de las cuestiones de referencia más comunes es la evaluación de los problemas de memoria y la diferenciación entre demencia y depresión, particularmente en los viejos (Butters, Delis y Lucas, 1995; Poon, 1986; Reeves y Wedding, 1994; Storandt y VandenBos, 1994). El cuadro diagnóstico se ve complicado además por los factores de la experiencia, que en algunos casos pueden no estar relacionados con el daño cerebral y que en otros pueden estarlo más o menos directamente. En consecuencia, para hacer la interpretación de cualquier signo diagnóstico específico de la ejecución en la prueba es necesario contar con información adicional sobre los antecedentes y el historial personal del individuo; por ejemplo, el nivel de habilidad del sujeto antes de la enfermedad es una información importante que se necesita para valorar la magnitud de los déficit cognoscitivos, así como el grado de recuperación de las funciones intelectuales (véase, por ejemplo, Matarazzo, 1990). A menudo, el nivel de escolaridad se usa como indicador amplio del funcionamiento premórbido, pero también se han desarrollado otros métodos de estimación que se basan en los datos históricos y el desempeño posterior al trauma en la prueba (Vanderploeg, 1994b).

En suma, la práctica de la evaluación neuropsicológica es una de las tareas clínicas más exigentes. Requiere la aplicación del conocimiento general sobre el funcionamiento cognoscitivo, de la personalidad, neurológico y fisiológico tanto en condiciones normales como patológicas. Por consecuencia, no resulta sorprendente que la bibliografía sobre la investigación en este campo y los manuales prácticos diseñados para orientar a estudiantes y profesionales continúe acumulándose a un ritmo prodigioso. El *Handbook of Neuropsychology*, compilado por Boller y Grafman (1988–1995), presenta en varios volúmenes el compendio de la investigación y la información clínica. Por otro lado, Adams, Parsons, Culbertson y Nixon (1996), Golden, Zillmer y Spiers (1992), Lezak

(1995), Touyz, Byrne y Gilandas (1994), Vanderploeg (1994a) y R. F. White (1992) han preparado recientemente trabajos de referencia sobre la práctica general de la evaluación y la intervención neuropsicológica, a la vez que continúan apareciendo otros tratados más especializados en temas como la neuropsicología forense o legal (Valciukas, 1995), la evaluación neuropsicológica de la exposición laboral a neurotóxicos (Agnew y Masten, 1994) y la evaluación neuropsicológica de hispanoparlantes (Ardila, Rosselli y Puente, 1994), entre muchos otros.

Instrumentos neuropsicológicos. Hay un número considerable de pruebas que han sido diseñadas como instrumentos clínicos para la evaluación de daños neuropsicológicos[42] (véase, por ejemplo, Lezak, 1995; Spreen y Strauss, 1991), que a menudo se designan como indicadores de la "organicidad" o del daño cerebral. Entre las principales funciones evaluadas por estos instrumentos se encuentran las que se consideran más sensibles a los procesos patológicos, como la percepción de las relaciones espaciales y la memoria para el material recién aprendido. El Test Gestáltico Visomotor de Bender, más conocido como el Test Gestáltico de Bender (*Bender-Gestalt Test*, BGT; Bender, 1938; Canter, 1996; Heaton, Baade y Johnson, 1978; Koppitz, 1964, 1975; Pascal y Suttell, 1951) y el Test de Retención Visual de Benton, quinta edición (*Benton Visual Retention Test-Fifth Edition*, BVRT, Sivan, 1991) ejemplifican estos instrumentos y durante muchas décadas se han utilizado como pruebas de identificación; sin embargo, dada la diversidad de disfunciones cerebrales orgánicas y los déficit conductuales acompañantes, ninguna prueba es adecuada para la identificación del daño cerebral en general, e instrumentos separados como el BGT y el BVRT son incluso menos adecuados para el diagnóstico diferencial.

Muchas veces, los neuropsicólogos clínicos utilizan una combinación de los instrumentos de que disponen para evaluar habilidades y deficiencias, en lo que ha llegado a conocerse como la aproximación de la "batería flexible" (véase, por ejemplo, Bauer, 1994; Goodglass, 1986), un procedimiento que tiene la ventaja de ofrecer combinaciones de pruebas adecuadas a los problemas que se presentan en cada caso, pero que también tiene graves limitaciones. Es probable que exista una duplicación innecesaria de funciones entre las pruebas y, por otro lado, que se descuiden algunas áreas importantes. La selección previa de los instrumentos adecuados para cada caso supone una carga pesada sobre la experiencia y el juicio del clínico; más aún, es poco probable que las pruebas desarrolladas independientemente sean comparables en cuanto a normas y escalas de calificación. También es probable que los datos empíricos sobre las relaciones de los distintos instrumentos sean magros, todo lo cual trae como consecuencia la dificultad de interpretar los resultados en términos de patrones de calificación.

Por estas razones, se han realizado esfuerzos sistemáticos por armar baterías estandarizadas que proporcionen medidas de todas las habilidades neuropsicológicas significativas. Dicha batería cumpliría diversas funciones: puede detectar con un alto grado de éxito el daño cerebral, también puede ayudar a identificar y localizar las áreas cerebrales lesionadas, diferenciar entre síndromes particulares asociados con las patologías cerebrales y colaborar en la planeación del entrenamiento de rehabilitación al revelar

[42] Las pruebas de esta categoría se revisan regularmente en los *Mental Measurements Yearbooks* (véase una lista actualizada en TIP-IV, p. 1116).

el tipo específico y el grado de los déficit conductuales. Dos ejemplos importantes de estas baterías neuropsicológicas son la Batería de Pruebas Neuropsicológicas de Halstead-Reitan (*Halstead-Reitan Neuropsychological Test Battery*, HRB; Reitan y Wolfson, 1993) y la Batería Neuropsicológica Luria-Nebraska (*Luria-Nebraska Neuropsychological Battery*, LNNB; Golden, Purish y Hammeke, 1985). Ambas baterías tienen versiones para niños y comparten el mismo propósito, pero difieren en varios aspectos importantes. La HRB, que es la más antigua, fue elaborada por Reitan a partir del trabajo de Halstead (1947) y ofrece a los examinadores cierta flexibilidad en el número y la selección de las pruebas que han de ser aplicadas (para revisiones, véase Dean, 1985; M. J. Meier, 1985).[43] La LNNB incorpora parte de la teoría y de los procedimientos diagnósticos de Luria (Christensen, 1975; Luria, 1973, 1980), tiene una mayor estandarización en su contenido, materiales, aplicación y calificación y su aplicación requiere considerablemente menos tiempo que la HRB (para revisiones, véase J. H. Snow, 1992; y Van Gorp, 1992).

Los adelantos recientes en la evaluación directa del daño cerebral mediante la electroencefalografía y las técnicas de imagenología nerviosa, como la imagenología por resonancia magnética (*magnetic resonance imaging*, MRI) y la tomografía por emisión de positrones (*positron emission tomography*, PET), tienen un profundo impacto sobre la neuropsicología básica y clínica. Aunque las tecnologías disponibles mejoran continuamente, ninguna técnica de diagnóstico es 100 por ciento confiable. En la mayor parte de los casos, los neuropsicólogos trabajan con los neurólogos y otros especialistas para obtener de diversas fuentes información corroborativa. En el contexto clínico, las pruebas conductuales cuidadosamente estandarizadas, junto con otros procedimientos, cumplen una importante función en la evaluación, planeación y supervisión de los programas de rehabilitación, y en el campo de la ciencia básica, la integración de las metodologías neuropsicológicas y de la imagenología nerviosa ofrecen algunas de las mayores promesas en el avance del conocimiento acerca de las relaciones entre cerebro y conducta (véase, por ejemplo, Gur y Gur, 1991, 1994).

Identificación de problemas específicos de aprendizaje

Desde los años setenta existe una ola continua de programas para el diagnóstico y remedio de los problemas de aprendizaje. Cada vez es mayor la conciencia de los educadores respecto a la elevada frecuencia de esta desventaja entre los escolares e incluso entre universitarios y otros adultos (véase, por ejemplo, Gregg, Hoy y Gay, 1996; Kravets y Wax, 1992; Wang, Reynolds y Walberg, 1991), aunque casi todos los individuos que se encuentran en esta categoría pueden reflejar en parte una mala clasificación que resulta de la vaguedad con la que se ha aplicado el término. La variada terminología aplicada a esos casos manifiesta tanto la modificación de las aproximaciones con el tiempo como las diferencias entre las orientaciones médica, educativa y psicológica a los problemas de aprendizaje (véase, por ejemplo, American Psychiatric Association, 1980, 1994). De

[43] Russell y Starkey (1993) elaboraron un sistema integrado basado en la computadora para compilar las puntuaciones no procesadas de muchas de las pruebas en la HRB (así como en otros instrumentos) y transformarlas en una escala uniforme, corregida por edad y nivel educativo. El programa también hace perfiles y resume los resultados para facilitar la interpretación. Heaton, Grant y Matthews (1991) prepararon un sistema similar en forma impresa que recientemente fue computarizado.

acuerdo con la definición establecida por el gobierno federal de los Estados Unidos en la Ley Pública 94–142, que proporciona educación a los niños con diversas deficiencias, (véase el capítulo 9) un *problema específico de aprendizaje* se describe como:

> . . . un trastorno en uno o más de los procesos psicológicos básicos involucrados en la comprensión o el uso del lenguaje, hablado o escrito, que puede manifestarse en una habilidad imperfecta para escuchar, pensar, hablar, escribir, deletrear o hacer operaciones matemáticas. El término incluye condiciones como las discapacidades perceptuales, el daño cerebral, la disfunción cerebral mínima, la dislexia y el desarrollo de la afasia. El término no abarca a los niños que muestran problemas de aprendizaje, que son principalmente resultado de discapacidades visuales, auditivas o motoras, de retardo mental, perturbación emocional o de desventajas ambientales, culturales o económicas (*Federal Register*, 1977, p. 65083).

La mencionada ley especifica además que el diagnóstico de un problema de aprendizaje sólo debe aplicarse a los niños que: (1) muestran una "discrepancia considerable" entre la capacidad intelectual y el rendimiento en diversas habilidades matemáticas y de comunicación, y (2) no muestran un aprovechamiento proporcional a su nivel de edad y habilidad, incluso cuando reciben educación apropiada.[44] De esta definición se deriva que no debería utilizarse la designación de los problemas de aprendizaje mientras no se hayan descartado otras condiciones como causas de las dificultades educativas o psicológicas del niño. Con los años, se ha reconocido que la heterogénea población con problemas de aprendizaje comprende subgrupos que pueden diferenciarse en términos de síntomas (Rourke, 1990; Feagans, Short y Meltzer, 1991; Geary, 1993; S. R. Hooper y Willis, 1989; Pennington, 1991; Shankweiler *et al.*, 1995; H. L. Swanson y Keogh, 1990). No obstante, incluso entre los especialistas en los problemas de aprendizaje persisten grandes diferencias en la orientación teórica, las que se reflejan tanto en los instrumentos de evaluación como en los programas de remedio.

Por lo general, los niños con problemas de aprendizaje muestran una inteligencia normal o superior a lo normal, en combinación con pronunciadas dificultades para aprender una o más de las habilidades educativas básicas (sobre todo la lectura); sin embargo, debe mencionarse que algunos problemas específicos de aprendizaje pueden ocurrir a cualquier nivel intelectual, aunque los niños con retardo mental que presentan problemas de aprendizaje no cumplen la definición legal de tales problemas. Entre los síntomas destacan las dificultades para percibir y codificar la información, una mala integración de las entradas de distintas modalidades sensoriales y la interrupción de la coordinación sensoriomotora. Es común que estos niños presenten perturbaciones en el desarrollo del lenguaje, en la memoria, el control de la atención y en las habilidades conceptuales, así como ciertos síntomas motivacionales y emocionales. En particular pueden desarrollar agresión y otros problemas afectivos y de trato personal, a menudo como resultado directo del fracaso académico y de las frustraciones generadas por las dificultades de aprendizaje. Al evaluar su conducta, debe tenerse en mente

[44] El lector puede encontrar en C. R. Reynolds (1990, 1992b) una revisión adicional de la P. L. 94–142 y sus ramificaciones sobre el diagnóstico de los problemas de aprendizaje. Por su parte, Sattler (1988, pp. 598–617) ofrece una revisión muy informativa de los problemas de aprendizaje y su evaluación.

que muchas dificultades específicas que son normales a una edad temprana (por ejemplo, a los tres años) representan una disfunción si persisten en edades mayores. En consecuencia, se necesita contar con un marco de referencia del desarrollo, con normas cualitativas, si no es que cuantitativas.

Técnicas de evaluación. Independientemente de la orientación teórica, existe el acuerdo general de que la identificación de los problemas de aprendizaje requiere una amplia variedad de instrumentos y procedimientos complementarios de observación. Esto se deriva de al menos tres características del problema del diagnóstico: (1) la variedad de trastornos conductuales asociados con esta condición; (2) las diferencias individuales en la combinación particular de síntomas, y (3) la necesidad de información específica respecto de la naturaleza y la extensión del problema en cada caso.

Por lo general, la evaluación de los niños con problemas de aprendizaje representa el esfuerzo conjunto de un equipo de profesionales. El maestro puede aplicar pruebas colectivas en el salón de clases y emplear otros instrumentos de identificación de banda ancha. Las baterías regulares de aprovechamiento también sirven a este respecto, en especial las diseñadas para los niveles de preescolar y primaria, y permiten análisis "referidos al criterio" (es decir, referidos al contenido) de las ventajas y desventajas específicas. Muchos de los instrumentos que citamos en la sección de evaluación escolar del capítulo son apropiados para ese propósito.

Las pruebas de aprovechamiento de amplio alcance, que se aplican individualmente, resultan muy útiles para la evaluación de los problemas de aprendizaje. Éstas pueden aplicarlas los maestros, aunque las observaciones cualitativas complementarias y la interpretación de las calificaciones pueden ser manejadas mejor por el clínico. El Test de Aprovechamiento Educativo de Kaufman (*Kaufman Test of Educational Achievement*, K-TEA; Kaufman y Kaufman, 1985) encaja en esta categoría. En la figura 17.4 se ilustra la aplicación de esta prueba, con el equipo tipo atril. Algunas baterías inclu-

Figura 17.4. Aplicación del Test de Aprovechamiento Educativo de Kaufman (K-TEA) que muestra el uso de un Equipo Atril. Las tarjetas en el aro del cuaderno forman un atril que se sostiene solo y que al mismo tiempo que muestra al examinado cada reactivo de la prueba le presenta al examinador las direcciones correspondientes.

(Cortesía del American Guidance Service.)

yen ciertos subtests clasificados como medidas de aptitud o de habilidad cognoscitiva y otros clasificados como medidas de aprovechamiento. Entre las más generales se encuentra la Batería Psicoeducativa de Woodcock-Johnson, Revisada (*Woodcock-Johnson Psycho-Educational Battery-Revised*, WJ-R; McGrew, 1994; McGrew, Werder y Woodcock, 1991). Instrumentos como las escalas de Stanford-Binet, la C-ABC y las de Wechsler no sólo proporcionan un indicador global, como el CI, para ayudar a diferenciar entre el retardo mental y los problemas específicos de aprendizaje, sino que también brindan mucha información cualitativa relativa a ciertas deficiencias (Kaufman, 1990, 1994; Sattler, 1988, 1992). Por ejemplo, estos instrumentos pueden revelar posibles deficiencias en la percepción y el recuerdo de patrones visuales, dificultades motoras, limitaciones de la memoria a corto plazo, incapacidad para manejar conceptos abstractos y muchos tipos de trastornos lingüísticos. Los editores de las escalas de Wechsler también desarrollaron el Test de Aprovechamiento Individual de Wechsler (*Wechsler Individual Achievement Test*, WIAT; Psychological Corporation, 1992), una batería comprensiva unida a las escalas de inteligencia que pretende complementarlas en la evaluación de los problemas de aprendizaje.

Aunque los instrumentos disponibles pueden ser de ayuda en la implantación de las directrices para la evaluación de los problemas de aprendizaje, muchos investigadores han exigido una nueva y más informativa aproximación al diagnóstico y evaluación de los niños con problemas de aprendizaje. Sus llamados se centran principalmente en el hecho de que, para mejorar tanto la evaluación como los esfuerzos de remedio es necesario contar con una base teórica clara que dirija la práctica de la evaluación y una comprensión más profunda de los procesos específicos que participan en cada caso (véase, por ejemplo, Das, Naglieri y Kirby, 1994; R. B. Kline, Snyder y Castellanos, 1996; C. R. Reynolds, 1992b).

Evaluación dinámica. El término "evaluación dinámica" cubre una variedad de procedimientos clínicos que comprenden en esencia el alejamiento deliberado de la aplicación de pruebas uniformes o estandarizadas para obtener datos cualitativos adicionales sobre el individuo. Aunque los clínicos experimentados han empleado estos procedimientos por algún tiempo, la aproximación se ha popularizado desde la década de los setenta (Lidz, 1981, 1987, 1991, 1995). Se ha empleado como fuente complementaria de datos, no sólo en los casos de problemas específicos de aprendizaje, sino también con otros niños que experimentan dificultades en la escuela, como aquellos con retardo ligero o moderado. También se ha explorado la utilidad de la aproximación para la evaluación de los sobredotados, en especial en el caso de los niños con desventajas económicas (véase, por ejemplo, Bolig y Day, 1993).

Una de las primeras adaptaciones cualitativas del procedimiento de evaluación se conoce como "evaluación de los límites", en la que el examinador puede proporcionar indicaciones adicionales. Entre más necesarias sean las indicaciones para un desempeño satisfactorio, mayor será el problema de aprendizaje. Las alteraciones del procedimiento estándar usadas para evaluar los límites son similares a algunas de las adaptaciones especiales introducidas al examinar a personas con discapacidades físicas, y las mismas advertencias y limitaciones interpretativas se aplican al desempeño resultante (véase el capítulo 9).

Hay una aproximación más reciente que se conoce como *evaluación del potencial para el aprendizaje* (Babad y Budoff, 1974; Campione y Brown, 1979; Feuerstein, 1979;

Glutting y McDermott, 1990; Hamers, Sijtsma y Ruijssenaars, 1993; Lidz, 1991). En el título, el término "potencial" tiene la connotación no garantizada de que la habilidad en cuestión está siempre presente y que sólo es necesario "descubrirla"; sin embargo, estos procedimientos incluyen un formato prueba-enseñanza-prueba en el que se enseña al estudiante por diversos medios a realizar la tarea que al principio no podía desempeñar. Superficialmente, la técnica es similar a la que se sigue en algunas de las pruebas de pronóstico educativo en que los examinados reciben una tarea de muestra que requiere el aprendizaje que adquirirán en un determinado curso. La presente aplicación difiere de las pruebas de pronóstico al menos en dos formas: (1) el examinado recibe instrucciones o sugerencias especiales y (2) las tareas generalmente incluyen habilidades de aprendizaje o para la resolución de problemas más ampliamente aplicables.

Las técnicas de evaluación dinámica iniciadas por Feuerstein y otros resultan prometedoras en varias direcciones. Al vincular la evaluación a la instrucción, promueven la investigación sobre la posibilidad de modificar la aptitud escolar y sobre el desarrollo de programas de remedio óptimos. Además, proporcionan al clínico calificado un instrumento de evaluación que produce descripciones más vívidas de la ejecución cognoscitiva y de su sensibilidad a las intervenciones de remedio que los tests estandarizados de inteligencia. A pesar de sus ventajas, las técnicas dinámicas no han escapado a las críticas. Una atañe a su *transportabilidad*, o el grado al que diversos clínicos pueden usarlas con eficacia; otra tiene que ver con la posibilidad de *generalizar* los efectos de remedio obtenidos con estas tareas tan amplias (por lo general los problemas de razonamiento no verbal similares a las matrices progresivas de Raven o los reactivos de la escala de ejecución de Wechsler) al desempeño escolar real. Más aún, aunque a muchos de los exponentes de las técnicas de la evaluación dinámica les gustaría alejarse de la confianza en las calificaciones que es característica de las medidas cognoscitivas "estáticas" (como los tests de inteligencia), el problema de cómo documentar el cambio sin usar números —de alguna manera— ha demostrado ser irresoluble (A. L. Brown *et al.*, 1992; R. E. Snow, 1990). No obstante, continúa la exploración de la utilidad de las técnicas de evaluación dinámica. Algunos de los trabajos prometedores presentados recientemente incluyen la aclaración de la función que cumplen diferentes estrategias de instrucción en la comprensión de conceptos matemáticos y el desarrollo de un sistema computarizado de evaluación dinámica para la ejecución en problemas de multiplicación con varios dígitos (Gerber, Semmel y Semmel, 1994; Jitendra, Kameenui y Carnine, 1994).

Desde otro ángulo, el modelo multidimensional de rasgo latente de Embretson (1987, 1990, 1992) evita muchas de las dificultades técnicas inherentes al uso tradicional de las pruebas para medir el cambio en la ejecución o el aprendizaje (véase, por ejemplo, Cronbach y Furby, 1970). La aproximación de Embretson utiliza la teoría de respuesta al reactivo (TRR, revisada en el capítulo 7) y la evaluación adaptada a la computadora (que vimos en el capítulo 10) para eludir esos problemas al evaluar a cada individuo con los reactivos cercanos a su nivel de umbral, en el que los efectos del aprendizaje son mayores; los reactivos en este nivel también proporcionan la estimación más confiable de la ejecución de cada individuo. Más aún, al aplicar los procedimientos de descomposición de tareas desarrollados en la psicología cognoscitiva, Embretson es capaz de variar sistemáticamente los procesos cognoscitivos específicos que se requieren para realizar la tarea presentada por cada reactivo de la prueba. Esto puede hacerse,

por ejemplo, utilizando subtareas que sólo requieran ciertos pasos de la solución completa, presentando indicios o conocimientos específicos requeridos por uno o más pasos en la solución o bien proporcionando instrucciones que influyan en aspectos particulares del desempeño de la tarea. Es evidente que la contribución de Embretson representa un logro importante en la psicometría que une los adelantos de campos tan diversos como la psicología clínica, la investigación cognoscitiva, las estadísticas y la tecnología de la computación (Embretson, 1993, 1995).

Evaluación conductual

Las diversas técnicas reunidas bajo el concepto general de modificación de conducta representan una utilización directa de los más importantes principios del aprendizaje al manejo práctico del cambio conductual. Básicamente, las técnicas incluyen la aplicación de los principios del condicionamiento a la adquisición y el fortalecimiento de las conductas deseadas y la eliminación de la conducta no deseada. La terapia conductual se ha ido ampliando para comprender una diversidad de problemas psicológicos, un repertorio creciente de técnicas de intervención y la consideración —además de las respuestas motoras abiertas— de las respuestas cognoscitivas y afectivas (Bandura, 1969, 1986; Hersen *et al.*, 1991; Lazarus, 1981).

Técnicas de evaluación. En las primeras aplicaciones de la terapia conductual se prestaba poca atención a la evaluación; sin embargo, desde mediados de los setenta cada vez es mayor el reconocimiento de la importancia de los procedimientos de evaluación (Barrios, 1988; A. S. Bellack y Hersen, 1988; Haynes, 1991; Mash y Terdal, 1988; Nelson y Hayes, 1986; O'Brien y Haynes, 1993; Ollendick y Hersen, 1993).[45] Las funciones principales que cumplen los procedimientos en la terapia conductual pueden resumirse en tres encabezados. Primero, las técnicas de evaluación ayudan a *definir el problema del individuo* mediante el análisis funcional de la conducta relevante. En esencia, dicho análisis incluye la especificación completa del objetivo del tratamiento, como superar una fobia o pensamientos obsesivos, e incluye la descripción de los estímulos que provocan el objetivo de la conducta, las situaciones en las que ocurre dicha conducta y la naturaleza, magnitud y frecuencia de las respuestas. La segunda forma en que los procedimientos de evaluación pueden guiar al terapeuta conductual es al *seleccionar los tratamientos apropiados*. En tercer lugar está la necesidad de *evaluar el cambio conductual* que resulta del tratamiento. Tales evaluaciones deberían incluir técnicas para supervisar el cambio, de modo que permitan evaluar la efectividad del tratamiento y, de ser necesario, la introducción de modificaciones, así como medidas terminales para establecer si se alcanzó un estado satisfactorio y planear los procedimientos necesarios de seguimiento.

Al considerar los procedimientos específicos de evaluación debemos advertir, en primer lugar, que el mismo procedimiento a menudo puede proporcionar información que es relevante para las tres funciones. Segundo, la elección de los procedi-

[45] Los instrumentos de esta categoría se revisan en los *Mental Measurements Yearbooks* (en TIP-IV pp. 1095–1096 encontrará una lista actualizada). En Kratchowill *et al.* (1991) puede hallar información sobre el uso de la tecnología computarizada en diversos procedimientos de evaluación conductual.

mientos depende de la naturaleza del problema, las características del cliente —incluido el ambiente en el que debe operar— y las facilidades de que se dispone en la clínica particular. Tercero, en casi todos los casos resulta deseable la combinación de varios procedimientos de evaluación.

Los procedimientos de evaluación disponibles pueden clasificarse a la vez en tres categorías principales: autodescripción del cliente, observación directa del cliente y medidas fisiológicas. Aunque no todo centro puede darse el lujo de contar con las instalaciones requeridas para obtener las *medidas fisiológicas*, éstas proporcionan datos objetivos complementarios en la evaluación de ciertas condiciones, como la ansiedad, los trastornos de excitación sexual y los trastornos del sueño (Sturgis y Gramling, 1988). Los ejemplos abarcan medidas de la actividad electrodérmica, muscular y electrooculár, así como del funcionamiento cardiovascular, sexual y cerebral.

La *observación directa* del objeto de la conducta puede realizarse en situaciones reales por padres, maestros, personal de la institución u observadores especializados; es posible utilizar varios apoyos de la observación como listas de verificación, escalas de valoración y programas diarios. Estas observaciones sufren de varias debilidades (Barrios, 1993; véase también el capítulo 16), por lo que a menudo se emplean situaciones análogas en las clínicas. También se explora el uso de sensores o monitores mecánicos que pueden proporcionar un registro objetivo continuo de la conducta en situaciones tanto reales como planeadas (W. W. Tryon, 1985, 1991).

La *autodescripción del cliente* comprende una diversidad de técnicas que incluyen entrevistas clínicas hechas por el terapeuta, registros de la conducta objetivo y condiciones asociadas realizados por el propio cliente y una variedad de inventarios y listas de verificación. Se utilizan algunos inventarios estandarizados autodescriptivos, ya sea en su versión original o en una adaptación, para la identificación preliminar, así como para la verificación de las conductas objetivos. Entre los más sencillos y más utilizados se encuentra la revisión del Inventario de Depresión de Beck (*Beck Depression Inventory*, BDI; Beck y Steer, 1993), un instrumento de 21 reactivos diseñado para evaluar la gravedad de la depresión mediante autovaloraciones.[46] Una de las herramientas estandarizadas más recientes es el Inventario de Consumo del Alcohol (*Alcohol Use Inventory*; Horn, Wanberg y Foster, 1990), un inventario autodescriptivo de 228 reactivos que evalúa la ingestión de alcohol de una manera conceptual y psicométricamente sofisticada que incluye escalas múltiples a diferentes niveles de generalidad (para revisiones, véase Drummond, 1995; McNeely, 1995). Muchos otros instrumentos se han preparado para emplearse en proyectos de investigación o en programas de tratamiento específicos, y aunque algunos no han sido publicados, por lo general se los describe y reproduce por completo en libros y artículos de revistas (véase, por ejemplo, A. S. Bellack y Hersen, 1988).

Recientemente se han elaborado otros instrumentos que comprenden escalas objetivas de valoración que son completadas por múltiples informantes. Una de las más amplias es el Sistema de Evaluación Conductual para Niños (*Behavior Assessment System for Children*, BASC) de Cecil Reynolds y Randy Kamphaus (1992), que estudiamos en el capítulo 16. El sistema incluye escalas de calificación de conducta para

[46] Véase en la decimoprimera edición de los MMY una extensa bibliografía sobre el BDI en las revisiones de una edición anterior.

padres y maestros y un forma para codificar y registrar las observaciones de la conducta en el salón de clases; además, proporciona un cuestionario de autorreporte para los mismos niños y un programa estructurado de entrevistas mediante el cual los padres pueden proporcionar la historia del desarrollo. El Sistema de Valoración de Habilidades Sociales (*Social Skills Rating System*, SSRS; Gresham y Elliot, 1990) también ofrece formas para los padres, los maestros y los mismos estudiantes para evaluar las conductas positivas y las problemáticas que presentan en los medios escolar y familiar. Según los revisores, uno de los aspectos útiles del SSRS es que proporciona un componente que permite unir los resultados de la evaluación con la planeación de las estrategias de intervención (para revisiones, véase Benes, 1995; Furlong y Karno, 1995).

Evaluación de las carreras

La práctica de la evaluación de las carreras supone ayudar al individuo a decidir cuál es la carrera más apropiada para él, teniendo en cuenta sus habilidades, intereses, metas, valores y temperamento, así como los requisitos de la ocupación. Pocas áreas de la vida son tan importantes para la gente como su ocupación, no sólo por la gran cantidad de tiempo que la mayoría dedica a trabajar, sino también porque el trabajo proporciona oportunidades para recibir muchas recompensas intrínsecas y extrínsecas (Super y Šverko, 1995). Además, el clima actual de rápidos cambios en la naturaleza y las condiciones del empleo está llevando a muchas personas a considerar su elección profesional no una sino muchas veces en el curso de la vida. Por ello no resulta sorprendente que la teorización en el campo de la elección y el desarrollo profesional proceda a un paso tan rápido. En los años noventa, el número de nuevas formulaciones teóricas importantes en el área igualó o superó a las producidas desde los cincuenta cuando Donald Super (1953) y John Holland (1959) hicieron sus primeras contribuciones importantes al campo.[47]

Desde el punto de vista de la evaluación, ya hemos considerado los instrumentos que mejor se aplican en la consejería profesional de los individuos, es decir, los inventarios de intereses, que explicamos en el capítulo 14, y las baterías de aptitudes múltiples, analizadas en el capítulo 10, así como en la sección de evaluación ocupacional de éste. La elección de una carrera a menudo implica la elección de un estilo de vida, con el conjunto de valores que lo caracteriza. Como los inventarios de intereses evalúan el sistema de valores del individuo, cada vez es mayor la importancia que se les concede en la planeación profesional. En esta sección examinaremos dos clases más especializadas de instrumentos que han sido diseñados especialmente para la consejería vocacional, programas generales para la exploración de carreras y medidas de la madurez vocacional. Muchos más instrumentos se describen y revisan en la indispensable guía de los instrumentos de evaluación profesional preparada por Kapes, Mastie y Whitfield, (1994), que ya se encuentra en su tercera edición.

[47] En Brown, Brooks *et al.* (1996) puede encontrar una revisión de ésta y otras teorías bien establecidas acerca de la elección de carrera, así como algunas contribuciones nuevas y provocativas. *Convergence in Career Development Theories*, de Savickas y Lent (1994), es otra fuente valiosa de información sobre las teorías y las formas en que se parecen, se distinguen y se complementan. Muchos de los capítulos de estos dos libros fueron escritos por los propios teóricos.

Programa comprensivo para la exploración de carreras. En los sistemas de orientación vocacional se han incorporado varias baterías de aptitudes múltiples. Un ejemplo es el Test de Aptitudes Diferenciales (*Differential Aptitude Tests*, DAT), que vimos en el capítulo 10 y que puede emplearse en combinación con el Inventario de Intereses Profesionales (*Career Interest Inventory*, CII; Psychological Corporation, 1991a, 1991b). Ambos instrumentos se desarrollaron conjuntamente para facilitar las comparaciones de sus resultados en el proceso de la orientación profesional.

Otro ejemplo es el programa elaborado por el Servicio de Empleo de los Estados Unidos, cuya Batería de Pruebas de Aptitudes Generales (GATB) analizamos en la sección de evaluación ocupacional del capítulo. Entre los instrumentos más útiles que han resultado del programa USES de consejo vocacional está la *Complete Guide for Occupational Exploration*, CGOE ("Guía Completa para la Exploración Ocupacional", Farr, 1992) y la *Enhanced Guide for Occupational Exploration*, EGOE ("Guía Mejorada para la Exploración Ocupacional", Maze y Mayall, 1995). Destinadas al uso de consejeros y de los propios estudiantes y de quienes buscan trabajo, las guías agrupan miles de ocupaciones en el mundo laboral de acuerdo con las principales áreas de interés, los patrones de habilidad y otros requisitos para el desempeño exitoso. El individuo puede usarlas para hacer una exploración preliminar de las carreras e identificar los trabajos en los que está más interesado para luego atender a la capacitación y las habilidades que requieren. La CGOE incluyen la lista de las 12 741 ocupaciones que aparecen en el *Dictionary of Occupational Titles* (U. S. Department of Labor, 1991), mientras que en la EGOE sólo se incluye 2 800 —lo que comprende el 95 por ciento de la fuerza de trabajo— pero proporciona más información sobre cada una.

Una nueva aproximación a la consejería profesional ofrece un procedimiento para integrar la información disponible de muchas fuentes en un programa comprensivo de exploración de carreras. La información puede incluir calificaciones de una gran variedad de pruebas (cada una de las cuales tiene sus propios datos normativos e interpretativos), datos biográficos (que incluyen educación y experiencia laboral) y los intereses, preferencias y sistema de valores del individuo. La aproximación la ilustran en diversos grados varios instrumentos disponibles, como el Sistema para la Toma de Decisiones Relacionadas con la Elección de Carrera, Revisado de Harrington-O'Shea, Revisado (Harrington y O'Shea, 1993) y el Programa de Planeación Profesional (*American College Testing*, ACT, 1994) del Programa Universitario de Evaluación.

Un sorprendente ejemplo de estos programas integrados es la versión revisada del Sistema Interactivo de Guía e Información (*System for Interactive Guidance Information*, SIGI-PLUS), que citamos en el capítulo 3. Usando un programa interactivo, SIGI-PLUS permite al individuo sostener una comunicación bidireccional con la computadora: cuestiona y responde preguntas, proporciona datos y solicita información. El programa incluye una amplia base de datos sobre las características y los requisitos de los trabajos y provisiones para incorporar otros datos locales. Aunque se diseñó para emplearse con estudiantes universitarios, el SIGI-PLUS ha sido actualizado para uso de adultos que desean hacer cambios profesionales o ingresar en el mercado laboral en diferentes etapas de la vida. El programa está diseñado para guiar al individuo en el examen de los hechos relevantes y proceder a lo largo de una ruta sistemática para tomar una decisión adecuada (M. R. Katz, 1993; Norris, Schott, Shatkin y Bennett, 1986). Sin embargo, ni siquiera semejante sistema diseñado para la aplica-

ción computarizada puede realizar por el individuo la función de decidirse por una carrera. Es necesaria la intervención de un consejero capacitado para estimular al individuo a pensar en sus necesidades y características particulares en diferentes etapas del desarrollo profesional (véase, por ejemplo, Tiedeman, 1994).

Evaluación de la madurez profesional. Hay otro instrumento destinado en particular a la consejería profesional que se interesa en el nivel de madurez profesional del individuo. El concepto surgió de un proyecto de investigación a largo plazo sobre el desarrollo profesional realizado por Super y sus colaboradores (Super *et al.*, 1957; Super y Overstreet, 1960). La *madurez profesional* significa al dominio que tiene el individuo de las tareas vocacionales apropiadas para su edad y a su eficacia para enfrentarlas. El estudio comprendió la investigación longitudinal de 20 años de cerca 100 muchachos de noveno grado. Los descubrimientos indican que, a nivel de la secundaria, la principal tarea de desarrollo profesional consiste en prepararse para tomar decisiones profesionales. Otra investigación sobre el desarrollo profesional, con aproximaciones transversales y longitudinales, ha estado aportando datos que contribuyen a redondear el cuadro (véase, por ejemplo, Crites, 1969; Gribbons y Lohnes, 1982; Super, 1980, 1985). Un descubrimiento importante de la investigación más reciente es que todas las conductas profesionales pueden ocurrir a lo largo de la mayor parte del ciclo vital, aunque predomine uno de ellos en cada etapa. Se está prestando mucha atención a los cambios de carrera en la vida adulta, cambios que obedecen a una diversidad de factores tanto en la vida personal como en la naturaleza de las ocupaciones (Kummerow, 1991; Lowman, 1991, 1993; Pickman, 1994; Walsh y Osipow, 1993).

Uno de los coproductos de la investigación sobre el desarrollo profesional es el establecimiento de medidas estandarizadas de la madurez profesional (Kapes *et al.*, 1994, pp. 241–272). Un ejemplo es el Inventario de Desarrollo Profesional (*Career Development Inventory*) preparado por Super y diseñado para evaluar la preparación para elegir una carrera y aclarar cualquier aspecto de la orientación profesional en la que los individuos necesiten ayuda (A. S. Thompson y Lindeman, 1981, 1984). Un instrumento relacionado es el Inventario de Creencias Profesionales (*Career Belifs Inventory*, Krumboltz, 1991), que se formuló como apoyo en la consejería con el propósito específico de identificar cualquier creencia que pudiera obstaculizar la obtención de las metas profesionales del individuo.

Juicio clínico

Lo que hace el clínico al evaluar al cliente puede considerarse un caso especial de cognición o percepción interpersonal, el proceso por el cual un individuo cualquiera llega a conocer y a comprender a otra persona (Kruglanski, 1989). Sin embargo, en la situación clínica la exactitud del juicio difiere de las circunstancias sociales comunes en varias formas significativas. Se ha escrito mucho sobre la función que cumple el clínico en el procesamiento, la síntesis o la interpretación de los datos. La investigación sobre el proceso del juicio clínico ha aclarado algunas de las fuentes de error, como la influencia de los estereotipos culturales y la confianza en principios falaces de predicción. Tal es el caso de la incapacidad para considerar los valores de bases o los efectos de regresión y la suposición de que entre mayor sea la correlación de los predictores más elevada será la validez (véase, por ejemplo, L. R. Goldberg, 1991).

Dado el mismo conjunto de hechos, como las puntuaciones de una prueba o los datos de la historia de vida, ¿proporcionará el juicio del clínico predicciones más exactas de la conducta subsecuente de las que se obtendrían de la aplicación rutinaria de una ecuación de regresión u otras fórmulas empíricamente derivadas? La pregunta tiene importancia teórica y práctica, ya que una vez que se ha desarrollado una estrategia estadística, o algoritmo, puede ser aplicada por un empleado o por una computadora para dar oportunidad al clínico de realizar otras funciones. En un libro clásico titulado *Clinical Versus Statistical Prediction*, Meehl (1954) hace un análisis del proceso de juicio clínico y estudia las investigaciones que comparaban ambas predicciones. Meehl demostró que, con una sola excepción cuestionable, la aplicación rutinaria de procedimientos estadísticos daba lugar al menos a tantas predicciones correctas como el análisis del clínico, y frecuentemente a más. La publicación del libro de Meehl y las confirmaciones subsecuentes de su conclusión estimularon una apasionada controversia que aún continúa activa (véase, por ejemplo, Anastasi, 1988b, pp. 511–515; Dawes, Faust y Meehl, 1993; Kleinmuntz, 1990).

A pesar de la aparente superioridad de la aproximación actuarial, cuando los procedimientos clínicos y estadísticos se aplican a los mismos datos es importante tener presente que el proceso clínico ofrece ciertas ventajas. Una contribución importante del método clínico, por ejemplo, es que mediante la entrevista cuidadosa y la observación de la conducta permite obtener datos en áreas en que no se dispone de instrumentos satisfactorios. El método clínico se adapta mejor que el estadístico al procesamiento de acontecimientos extraños e idiosincráticos cuya frecuencia es demasiado baja para permitir el establecimiento de estrategias estadísticas.

En resumen, para el estado actual de nuestro conocimiento, el procedimiento más eficaz combina las aproximaciones clínica y estadística (Matarazzo, 1990). Los clínicos deberían utilizar todos los datos objetivos y estrategias actuariales que puedan aplicarse a la situación particular y complementar al mismo tiempo la información con hechos e inferencias que sólo pueden lograrse por métodos clínicos. Siempre que sea factible, deben investigar sistemáticamente la validez de las predicciones clínicas comparada con los resultados.[48] Es necesario contar con más datos sobre la consistencia de las predicciones sobre las mismas personas hechas por diferentes clínicos y por los mismos clínicos en momentos distintos. En la medida de lo posible, en los registros clínicos deben hacerse explícitos el proceso y los indicios que fundan las predicciones clínicas. Esta práctica no sólo facilitaría la investigación y la formación, sino que también ayudaría a fomentar la confianza en datos adecuados e interpretaciones defendibles. Por último, el "clínico como instrumento" es un concepto importante, lo que se ilustra en la investigación en marcha sobre las características que moderan la precisión del juicio clínico. Mediante una serie de investigaciones en curso, Spengler y sus colegas han llegado a la conclusión de que los psicólogos con menores niveles de complejidad cognoscitiva tienen mayor probabilidad de formar juicios clínicos sesgados que aquellos con mayor complejidad (Spengler y Strohmer, 1994; Walker y Spengler, 1995).

[48] Actualmente se planean algunos proyectos a gran escala de esta naturaleza con el objeto de evaluar la validez de las decisiones o las predicciones basadas en los procedimientos de evaluación. Una de esas investigaciones, ya en proceso bajo los auspicios de la Sociedad para la Evaluación de la Personalidad (*Society for Personality Assessment*), es una revisión de toda la bibliografía —que ha de ser seguida por un meta-análisis— de los estudios que han empleado procedimientos de evaluación de la personalidad o instrumentos para hacer predicciones acerca de aspectos de la terapia o de los resultados médicos, legales o vitales (Handler y Meyer, 1996, Primavera/Verano).

El informe de evaluación: síntesis final. En el capítulo 18 examinaremos algunos de los temas más generales que atañen a la comunicación de los resultados de la prueba haciendo particular referencia a sus implicaciones éticas y sociales. Para el clínico, dicha comunicación generalmente incluye la preparación de un informe escrito de la prueba o del caso, que a menudo es seguido por el análisis o la consulta con el cliente, los padres, los maestros u otros profesionales. Incluso en las situaciones en las que no se requiere de un informe escrito, es buena idea preparar uno como registro para referencias futuras. La preparación del informe también ayuda a que el clínico organice y aclare sus ideas sobre el caso y agudice sus interpretaciones. La redacción del informe representa la etapa final en la función sintetizadora del clínico. En su contenido, el informe debe hacer uso de todas las fuentes de datos (pruebas y no pruebas) de que dispone.

Varios libros ofrecen orientación para la redacción del informe.[49] Sin duplicar las muchas listas de consejos que pueden encontrarse en esas fuentes, nos concentraremos en algunos de los puntos más importantes. Primero, no existe una forma estándar de esbozar todos los informes. Tanto el contenido como el estilo deben variar según el propósito de la evaluación, su contexto, las personas a las que se dirige el informe y la orientación clínica y los antecedentes profesionales del clínico. Resulta particularmente importante adaptar el informe a las necesidades, intereses y antecedentes de quienes lo reciben; por ejemplo, un informe dirigido a un abogado debe ser muy diferente del que se destina al psicoterapeuta; no obstante, en ambos casos el clínico debe seleccionar lo que sea relevante para dar respuesta a las preguntas planteadas desde el inicio de la gran cantidad de datos que ha obtenido.

El informe también debe concentrarse en las características que distinguen a cada individuo —los puntos altos y bajos— en lugar de los rasgos en los que su posición es cercana al promedio. Una prueba de la eficacia del informe consiste en ver si es único para el individuo o si se aplica igualmente bien a otras personas. Preparar un pseudoinforme a partir de las afirmaciones estereotipadas generales que se aplican a la mayoría de la gente resulta una tarea relativamente sencilla; se cuenta con un cuerpo considerable de investigación que demuestra que casi todas las personas aceptan con facilidad estos informes como autodescripciones "notablemente exactas" (Goodyear, 1990; Klopfer, 1983; Snyder y Larson, 1972; Tallent, 1992, pp. 236–238). Esta pseudovalidación se conoce como "efecto Barnum", en honor de Phineas T. Barnum, el famoso comediante al que se atribuye la frase que afirma que "cada minuto nace un incauto". La confianza en semejantes descripciones de la personalidad que pueden aplicarse de manera general es el recurso favorito de los adivinadores de fortuna y otros charlatanes.

La parte principal del informe debe estar en las interpretaciones y en las conclusiones, aunque en algunos casos es posible anexar por separado los registros de las pruebas y otros datos detallados.[50] Por lo general, sólo deben citarse datos concretos, como las respuestas o las puntuaciones de los subtests, para ilustrar o aclarar un punto. Los informes deben organizarse e integrarse cuidadosamente, deben escribirse de manera

---

[49] Sattler (1988, capítulo 23) ofrece una revisión general de los problemas y errores, con muchos consejos prácticos, y resume la investigación relacionada con la redacción de informes. Otras fuentes, con informes de casos ilustrativos, son Ownby (1991) y Tallent (1993).

[50] Véase el número de junio de 1996 de *American Psychologist* para una "declaración sobre la revelación de los datos de las pruebas" preparada por la Comisión de Evaluación y Pruebas Psicológicas de la Asociación Psicológica Estadounidense.

sencilla para comunicar y no para confundir. Los textos sobre la preparación de los informes de la evaluación suelen contener consejos para la buena redacción así como referencias a manuales de estilo. Un texto pequeño que además es divertido y que logra hacer la redacción menos dolorosa tanto para el escritor como para el lector es *The Elements of Style*, de Strunk y White (1979).

La función de las computadoras en la evaluación psicológica

Durante varias décadas se ha tenido acceso a la calificación computarizada de muchas clases de pruebas, incluyendo los inventarios autodescriptivos de personalidad. La mayor parte de los programas de computadora también realizan análisis estadísticos rutinarios, con lo que ofrecen mayor información sobre diversas puntuaciones derivadas, intervalos de calificación en términos del EEM y perfiles de calificación. Los informes narrativos constituyen una aplicación más compleja de la tecnología de la computación y la utilización de los datos de las pruebas (Butcher, 1987; Moreland, 1992). En esencia, estos programas utilizan una enorme base de datos de afirmaciones interpretativas cualitativas que se vinculan a determinados niveles o patrones de calificaciones cuantitativas. Además de ahorrar tiempo al clínico, tienen otras ventajas: la computadora puede buscar de manera sistemática y constante en una base de datos mucho más extensa de lo que podría hacer el clínico; también puede aplicar confiablemente reglas más complejas a los datos actuariales al seleccionar las declaraciones apropiadas y puede insertar en el proceso muchas otras variables relacionadas, como los datos demográficos de diferentes poblaciones normativas.[51]

La contribución de las computadoras a la evaluación psicológica es sin lugar a dudas impresionante (Butcher, 1987; Eyde, 1987; Gutkin y Wise, 1991; Moreland, 1992), si bien una gran parte de su potencial apenas comienza a explorarse (Embretson, 1992); por ejemplo, la aplicación de las técnicas de ramificación y de la evaluación adaptada, cuyas ventajas son ahora reconocidas en la evaluación de la habilidad, apenas empieza a incluirse en la evaluación de la personalidad (Ben-Porath y Butcher, 1986; Jackson, 1985, 1991) y todavía no se producen instrumentos adecuados para el uso clínico.

Por otro lado, la tecnología de la computadora ha llevado al desarrollo y la rápida proliferación de muchos instrumentos nuevos para la evaluación del funcionamiento cognoscitivo que empiezan a utilizarse en la neuropsicología clínica así como en la evaluación de problemas de aprendizaje y trastornos de atención (véase, por ejemplo, Krug, 1988, 1993; Stoloff y Couch, 1992). En el futuro previsible es poco probable que la evaluación adecuada del funcionamiento neurocognoscitivo se realice únicamente por medio de la aplicación computarizada de los instrumentos (Golden, 1987); no obstante, las computadoras hacen posible manipular de manera más precisa la presentación de las condiciones de manera que sea posible evaluar la ejecución en

[51] Casi todos los editores y distribuidores de programas de cómputo para los informes de perfiles de pruebas, informes narrativos y otras opciones para la interpretación de pruebas asistidos por computadoras ofrecen muestras y discos de demostración para los usuarios. Además, el *Psychware Sourcebook* (Krug, 1993), que es una de las fuentes de información más completas actualmente disponibles sobre los productos de evaluación basados en las computadoras, contiene reproducciones de muchos informes de muestra.

diferentes componentes de la tarea. También permiten el registro y la evaluación de los parámetros de respuesta —como el tiempo— de formas que no son posibles con las pruebas de lápiz y papel ni con las individualizadas. La Evaluación del Funcionamiento Cognoscitivo: MicroCog (MicroCog, Powell *et al.*, 1993; Powell y Whitla, 1994a, 1994b) es un ejemplo de una batería aplicada por computadora para identificar posibles signos de daño cognoscitivo en los adultos. MicroCog consta de 18 subtests —en las áreas de atención y control mental, memoria, razonamiento y cálculo, procesamiento espacial y tiempo de reacción— que hacen uso de las capacidades de la tecnología de las computadoras. Los beneficios de esta batería relativamente rápida y poco costosa incluyen la evaluación de niveles leves a moderados de deterioro cognoscitivo en adultos mayores y en la supervisión del desempeño cognoscitivo entre empleados que pueden estar expuestos a estímulos nocivos, así como otros casos que requieren la evaluación precisa de los cambios neurocognoscitivos.

También se han preparado muchos otros instrumentos nuevos para la evaluación de funciones específicas e indudablemente continuarán apareciendo en los próximos años. Entre los primeros instrumentos de esta naturaleza se encuentran varias pruebas de desempeño continuo como el Test de Variables de Atención (*Test of Variables of Attention*, TOVA; Leark, Dupuy, Greenberg, Corman y Kindschi, 1996), que está disponible tanto en versión auditiva como visual. Otros ejemplos incluyen el Test de Desempeño en Vigilia Continua (*Vigil Continuous Performance Test;* Cegalis, Cegalis y Bowlin, 1993) y el Test de Pasos Seriados de Adición Auditiva (*Paced Auditory Serial Addition Test*, PASAT; Cegalis y Birdsall, 1995), los cuales requieren paquetes de cómputo con multimedios que permiten la evaluación de la atención en las modalidades visual y auditiva. La figura 17.5 muestra a un niño que presenta la versión visual del TOVA.

Desde otro punto de vista, las computadoras también tienen el potencial para integrar los datos de muchas fuentes, incluyendo todas las clases de instrumentos disponibles, historias de casos y observaciones conductuales directas (véase, por ejemplo, Watkins y McDermott, 1991). Aunque la computadora podría asumir, por consecuencia, la función sintetizadora del clínico para casos individuales, aún no se dispone de la base de datos computarizada que es necesaria para elaborar y mantener estos programas integradores.

Entre las aplicaciones actuales también existen errores graves que deben ser considerados (Moreland, 1992). La mayor parte de los sistemas disponibles de interpretación de pruebas asistidas mediante computadora combina procedimientos clínicos y estadísticos. La combinación de los datos cuantitativos y el juicio clínico varía entre los sistemas, como también la calidad técnica de la base de datos y la experiencia clínica de los juicios. Más aún, la información necesaria para evaluar los sistemas a menudo no está disponible por cuestiones de patente.

Es en gran medida por la ausencia de la información técnica necesaria que los abusos potenciales de los programas de interpretación computarizada de las pruebas han generado tanta preocupación (Eyde y Kowal, 1987; Fowler y Butcher, 1986; Matarazzo, 1986a, 1986b; Moreland, 1987). Algunos programas ya cumplieron las especificaciones científicas y profesionales adecuadas o están siendo revisados para hacerlo; pero de varios se desconoce la calidad y nunca han sido revisados por profesionales calificados, y

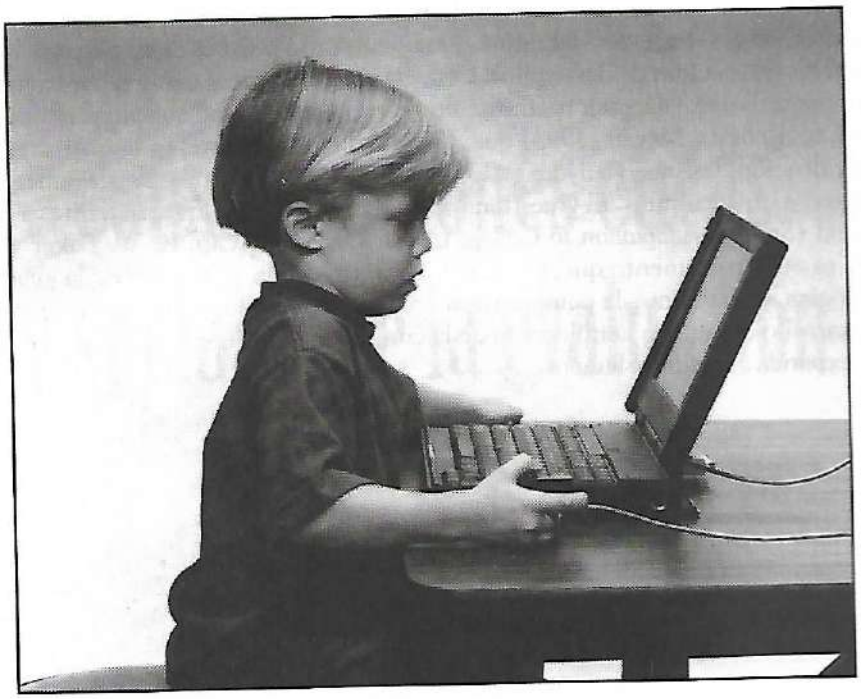


Figura 17.5. Un niño que presenta el Test de Variables de Atención (TOVA) responde pulsando un interruptor externo unido a uno de los puertos de la computadora.

(Cortesía del American Guidance Service.)

muchos se venden con afirmaciones no verificadas en los textos promocionales. Las primeras directrices para la evaluación y el uso de los servicios interpretativos computarizados fueron publicadas por la Asociación Psicológica Estadounidense en 1986; y en los nuevos *Estándares de la Examinación*, actualmente en preparación, se incorporan nuevas adiciones y revisiones de estas directrices. En algunas otras publicaciones se dispone de orientaciones adicionales sobre el uso de los instrumentos computarizadas para la evaluación psicológica (véase, por ejemplo, Bersoff y Hofer, 1991; Moreland, 1992).

Comentarios finales. En conjunto, el campo de la evaluación psicológica practicada por diversos especialistas está cambiando tan rápidamente como otras áreas que consideramos en el capítulo. Además de la gran cantidad de desarrollos basados en el uso de computadoras que continuamente tienen lugar y de las otras tendencias que presentamos en esta última sección, hay un énfasis renovado en la necesidad de contar con instrumentos de evaluación que estén orientadas hacia la salud mental más que hacia las psicopatologías. Un ejemplo reciente es el Inventario de la Calidad de Vida (*Quality of Life Inventory*, QOLI; Frisch, 1994), que es una medida de la satisfac-

ción con la vida que puede usarse en la planeación del tratamiento así como en la evaluación de los resultados del mismo (y muchas otras pruebas de este tipo se encuentran en diversas etapas de desarrollo). Otra clase interesante y única de instrumento es el Inventario de Adaptación Transcultural (*Cross-Cultural Adaptability Inventory*, CCAI; C. Kelley y Meyers, 1993) que, como indica su nombre, es un instrumento de autoevaluación destinado a ayudar a los individuos a decidir sobre su preparación para adaptarse a otras culturas. El Cuestionario de Adaptación de los Estudiantes a la Universidad (*Student Adaptation to College Questionnaire*, SACQ; R. W. Baker y Siryk, 1989) es otro instrumento que, como el CCAI, tipifica la aplicación de la evaluación psicológica a la mejora y la comprensión propias del individuo, una aplicación que es un desarrollo directo de la influencia de la consejería y que es probable que en el futuro se expanda considerablemente.

Consideraciones éticas y sociales de la evaluación

Los psicólogos se han preocupado por aspectos de la ética profesional en todo lo que concierne a su investigación y a las aplicaciones prácticas de sus procedimientos. Un ejemplo concreto es el programa sistemático empírico que se inició en la década de los cincuenta para establecer el primer código formal de ética para la profesión. Esta ambiciosa empresa dio por resultado la preparación de un conjunto de estándares que la APA adoptó oficialmente y publicó por primera vez en 1953. Estas normas han sido sometidas a continuas revisiones y perfeccionamientos, lo que ha dado lugar a la publicación periódica de ediciones revisadas. La versión actual, *Ethical Principles of Psychologists and Code of Conduct* ("Principios éticos y código de conducta de los psicólogos", APA, 1992),[1] comprende un preámbulo y seis principios generales destinados a orientar a los psicólogos hacia los *más altos ideales* de la profesión. Además, incluye ocho normas éticas con *reglas cuyo cumplimiento puede exigirse* a psicólogos que trabajan en diversos contextos.

El *Código Ético* fue puesto en práctica por La Comisión de Ética de la APA, que se encarga de investigar y juzgar las quejas contra los miembros de la asociación. Las reglas y los procedimientos para la operación de la comisión, así como sus primeros informes, se publicaron en el *American Psychologist*, la publicación oficial de la Asociación Estadounidense de Psicología (véase, por ejemplo, APA, Ethics Committee, 1995, 1996). Los miembros de la Comisión de Ética anterior, que participaron en la más reciente revisión del Código Ético, prepararon una guía indispensable en la que

[1] En un esfuerzo por difundir tan ampliamente como sea posible estos principios —a los que nos referiremos de ahora en adelante como *Código Ético*— la APA enviará una copia gratis a quien lo solicite. *Véase* en el Apéndice B la dirección de la APA.

pueden encontrarse comentarios y ejemplos de las aplicaciones de los criterios éticos (Canter, Bennett, Jones y Nagy, 1994). Bersoff (1995) presentó recientemente una compilación de lecturas históricas y contemporáneas sobre la ética en la psicología, incluidos algunos debates sobre dilemas éticos en diversos medios, que resulta de gran utilidad. Por otra parte, el volumen editado por Pope y Vásquez (1991) sobre la *ética en la psicoterapia y la consejería* incluye un capítulo sobre los temas de la evaluación en el que el lector puede encontrar muchos consejos prácticos; además, Weiner (1995a) escribió recientemente un capítulo que sin duda resultará de gran ayuda sobre la forma de anticipar los desafíos éticos y legales en la evaluación de la personalidad.

En los Estados Unidos, la década de los noventa ha presenciado una rápida proliferación de acciones legislativas, tanto a nivel federal como estatal, resoluciones de los tribunales y orientaciones profesionales desde diversos puntos de vista, muchas de las cuales tienen impacto en la práctica de la psicología en general y en el uso de las pruebas psicológicas en particular. En los capítulos anteriores vimos ya algunas de estas acciones en el contexto de temas específicos y prácticas de evaluación, cuyo efecto combinado ha generado, con demasiada frecuencia, confusión, incongruencias y conflictos entre los usuarios.

Cada vez es más común que se concentre la atención en las personas que proporcionan servicios psicológicos, quienes deben equilibrar los principios éticos de su profesión con los mandatos legales y los reglamentos así como también con las políticas institucionales de las organizaciones en las que trabajan. De acuerdo con ello, la APA se ha esforzado en proporcionar a sus miembros orientación e información, supervisar los adelantos pertinentes y promulgar normas, directrices y declaraciones sobre los temas que quizá presenten problemas en la práctica profesional. Las Directrices Generales para los Prestadores de Servicios Psicológicos (*General Guidelines for Providers of Psychological Services;* APA, 1987a) y las Directrices de especialidades para la entrega de servicios (*Specialty Guidelines for the Delivery of Services;* APA, 1981), actualmente en revisión, se promulgaron para ayudar a quienes participan en diversos contextos en la práctica profesional de la psicología. En otros documentos se ofrece orientación adicional sobre temas específicos, como las Directrices para las Evaluaciones del Niño en los Procedimientos de Divorcio (*Guidelines for Child Custody Evaluations in Divorce Proceedings*), un trabajo preparado por la Comisión de Prácticas y Normas Profesionales (APA, COPPS, 1994), y muchos otros que mencionaremos a lo largo del capítulo. Además, desde finales de los ochenta la APA prepara una serie de volúmenes que resumen y analizan las leyes que afectan a los profesionales de la salud mental en los Estados Unidos. Mientras preparábamos este texto, se había publicado alrededor de una docena de volúmenes y, en algunos casos, ya han sido actualizados (véase, por ejemplo, Caudill y Pope, 1995; Petrila y Otto, 1995; Shuman, 1990, 1993; Wulach, 1991).

La Comisión sobre Tests Psicológicos y Evaluación de la APA (*Committee on Psychological Tests and Assessment*, CPTA) se dedica en particular a la consideración de problemas relacionados con las prácticas adecuadas de examinación y evaluación y a proporcionar consejo técnico a otros grupos de la asociación respecto a esas prácticas. La CPTA se encarga también de revisar el trabajo de la Comisión Conjunta sobre Prácticas de Evaluación (*Joint Committee on Testing Practices*, JCTP), un grupo establecido por la APA y otras organizaciones profesionales interesadas en la evaluación, y que a su vez redactó el Código de Prácticas Justas de Evaluación en la Educación

(*Code of Fair Testing Practices in Education*; JCTP, 1988) y otros trabajos cuyo propósito es mejorar la forma en que se aplican las pruebas y prevenir su abuso (véase, por ejemplo, Eyde *et al.*, 1988, 1993). Actualmente, la JCTP prepara una declaración sobre los "Derechos y responsabilidades de los examinados".

En el primer capítulo consideramos el punto de vista del examinado y ciertos aspectos de la función de los usuarios de las pruebas; en éste, dirigiremos la atención a los aspectos éticos y sociales que influyen en el uso de las pruebas en diversos contextos. Además, para los temas que atañen a la competencia profesional, analizaremos brevemente las responsabilidades de las editoriales, el derecho del examinado a la privacía, el tema de la confidencialidad y la evaluación de personas con diversos antecedentes y niveles de capacidad. Aunque ya antes se ha hablado del impacto de la legislación, la consideración detallada de la multiplicidad de aspectos legales que afectan la práctica evaluativa va más allá de nuestro propósito, por lo que se recomienda al lector interesado que consulte las diversas fuentes mencionadas en éste y en otros capítulos del texto (especialmente, los capítulos 9 y 17).

ASPECTOS ÉTICOS EN LA EVALUACIÓN Y EXAMINACIÓN PSICOLÓGICA

Desde la década de los setenta ha aumentado la preocupación no sólo por los problemas éticos, sino por otras cuestiones relacionadas con los valores en todos los campos de la psicología teórica y aplicada (Bersoff, 1995; Diener y Crandall, 1978; Jacob y Hartshorne, 1991; Pope y Vásquez, 1991). En el área de la evaluación, Eyde y Quaintance (1988) y Messick (1980b, 1989, 1995) han presentado análisis minuciosos y provocativos sobre la función de los valores y de la ética racional que fundamenta diversas prácticas. A nivel más específico, el Código Ético de la APA contiene muchos elementos que pueden aplicarse a la evaluación psicológica. Una de las normas —de evaluación e intervención— contiene una sección dedicada específicamente a las evaluaciones en contextos legales; además, la norma ética sobre la privacía y la confidencialidad, aunque de gran alcance, también resulta relevante para la evaluación, como lo son la mayor parte de los otros principios generales y varias de las normas éticas (APA, 1992). Algunos de los asuntos analizados en el Código Ético se relacionan con los puntos cubiertos en los *Estándares de Examinación* que citamos en el primer capítulo, cuyo contenido de hecho ayuda a definir el uso responsable y profesional de las pruebas.

Además de la APA, otros grupos y asociaciones profesionales afines han desarrollado sus propios códigos y normas éticos, entre los que destacan, desde el punto de vista de la evaluación, la declaración titulada *Responsibilities of Users of Standardized Tests* ("Responsabilidades de los Usuarios de las Pruebas Estandarizadas"), la "declaración RUST" adoptada en 1989 por la ACA. Otro documento útil son los *Principles for the Validation and Use of Personnel Selection Procedures* ("Principios para la Validación y Uso de Procedimientos de Selección de Personal"), redactado por la Sociedad para la Psicología Industrial y Organizacional (*Society for Industrial and Organizational Psychology*, SIOP, 1987) para un propósito más especializado (véase el capítulo 17).

La publicación de *Ability Testing: Uses, Consequences, and Controversies* ("Evaluación de Habilidades: Usos, Consecuencias y Controversias", Wigdor y Garner, 1982) constituyó un acontecimiento importante para aclarar el lugar de la evaluación en la

sociedad moderna; la obra, que consta de dos volúmenes, es el informe final de un proyecto que durante cuatro años examinó el uso de pruebas estandarizadas de habilidad en la escuela, la admisión a la educación superior y el empleo. El proyecto, que empezó cuando el debate público sobre lo conveniente de la evaluación era generalizado, fue dirigido por una comisión multidisciplinaria bajo los auspicios del Consejo Nacional Estadounidense para la Investigación. Desde la década de los ochenta se han publicado otros estudios e informes importantes que tienen que ver con las áreas problemáticas de la evaluación (véase, por ejemplo, Hartigan y Wigdor, 1989; Office of Technology Assessment, 1992). En general, los descubrimientos de esos diversos grupos han apoyado y aumentado la credibilidad de las bien establecidas y a menudo reiteradas conclusiones sobre las contribuciones y los abusos potenciales de las pruebas de habilidad.

La creciente participación del gobierno estadounidense en la aplicación de las pruebas psicológicas y otras herramientas de evaluación dio lugar a la creación de la Comisión para la Examinación y Evaluación (*Board on Testing and Assessment*, BoTA), establecida en 1993 con el apoyo de los departamentos de Defensa, Educación y Trabajo de los Estados Unidos (véase el apéndice B). La BoTA supone una actividad en marcha del Consejo Nacional Estadounidense para la Investigación, que tiene entre sus objetivos principales ayudar a los políticos a comprender y evaluar las pruebas y otros instrumentos de evaluación que se emplean como herramientas de política pública. La comisión se concentra en los temas que influyen en la examinación y evaluación en diversos contextos y ha publicado informes al respecto como las *Metas del 2000*, una iniciativa legislativa para la educación (Feuer y Kober, 1995) y un plan para mejorar la Batería de Pruebas de Aptitudes Generales (BoTA, 1995), así como las implicaciones evaluativas y educacionales de los cambios que tienen lugar en la naturaleza del trabajo (Black, Feuer, Guidroz y Lesgold, 1996).

CAPACIDADES Y COMPETENCIA PROFESIONAL DEL USUARIO

El principio del *Código Ético* correspondiente a la competencia afirma que los psicólogos "proporcionarán sólo aquellos servicios y utilizarán las técnicas para las que estén calificados ya sea por educación, capacitación o experiencia" (APA, 1992, p. 1599). En lo que respecta a las pruebas, el requisito de que sólo deben ser empleadas por examinadores adecuadamente calificados pretende proteger al examinado contra el uso inadecuado de tales instrumentos.[2] Por supuesto, la capacitación necesariamente varía con el tipo de instrumento, de ahí que suela requerirse un periodo más bien largo de capacitación intensiva y práctica supervisada para el buen uso de los tests individuales de inteligencia y más para los de personalidad, mientras que en el caso de las pruebas de aprovechamiento educativo o eficiencia laboral se requiere una formación mucho menos especializada. Debe advertirse que los estudiantes que ocupan pruebas en el salón de clases con propósitos de enseñanza no están preparados para aplicarlas a otros o para hacer una interpretación adecuada de las puntuaciones.

[2] Véase el capítulo 1 para una revisión de la función del usuario y de los proyectos sobre la capacitación y la formación del usuario de las pruebas, conducidos por grupos de trabajo del JCTP (Eyde *et al.*, 1988, 1993; Moreland *et al.*, 1995). La Asociación Psicológica de Canadá y la Sociedad Psicológica Británica también han dado pasos hacia el desarrollo de sistemas para establecer la capacitación de los usuarios (D. C. Browns, 1995).

Los evaluadores bien capacitados eligen los instrumentos apropiados tanto para el propósito particular de evaluación como para las personas que serán examinadas. Están al tanto de la investigación bibliográfica disponible sobre la prueba elegida y pueden evaluar sus méritos técnicos con respecto a características como las normas, la confiabilidad y la validez. Al aplicar el instrumento, son sensibles a las diversas condiciones que pueden afectar el desempeño en el mismo, como las que se ilustran en el capítulo 1, extraen conclusiones o hacen recomendaciones únicamente después de considerar la puntuación (o puntuaciones) de la prueba a la luz de otra información relevante sobre el individuo. Por encima de todo, deben contar los conocimientos suficientes sobre la ciencia del comportamiento humano para prevenir las inferencias injustificadas en su interpretación de las puntuaciones. Cuando técnicos o asistentes aplican las pruebas, o bien personas que carecen de formación profesional en los principios psicométricos y las prácticas adecuadas de evaluación, resulta indispensable que tengan acceso a un psicólogo calificado, al menos como asesor, que les proporcione la perspectiva necesaria para hacer una interpretación adecuada de la ejecución en la prueba.

¿Quién es un psicólogo calificado? Como es evidente, y dada la diversificación de la disciplina y la consecuente especialización de la formación, no todos los psicólogos están igualmente calificados en todas las áreas, ni siquiera en el campo más restringido de las pruebas y la evaluación psicológica (véase el capítulo 17). En reconocimiento de este hecho, el *Código Ético* exige a los psicólogos que "acepten los límites de su competencia particular y las limitaciones de su experiencia" (APA, 1992, pág. 1599). Las implicaciones de esta obligación ética se explican claramente en el principio de competencia que mencionamos.

La promulgación de leyes para el otorgamiento de licencias estatales y de las leyes de certificación para los psicólogos no sólo contribuyeron al mejoramiento de los estándares profesionales, sino que también ayudan al público a identificar a los psicólogos calificados. Estas leyes tienen vigencia en los Estados Unidos, y en casi todas las provincias del Canadá también se han promulgado leyes que regulan la práctica de la psicología (el lector puede encontrar un resumen de esas leyes en APA, 1993, pp. *xliii-xlv*). Aunque es común el uso indistinto de los términos "licencia" y "certificación", en psicología la certificación suele referirse a la protección legal del título de "psicólogo", mientras que la licencia controla la práctica de la psicología independientemente del título que identifique al profesional. En consecuencia, las leyes sobre la licencia necesitan incluir una definición de la práctica de la psicología. Por su parte, en las leyes de certificación, aunque al principio eran casi siempre simples, hoy se observa una tendencia continua hacia la regulación de la licencia. Cualquiera que sea la ley, por lo común se exige un doctorado en psicología, cierta cantidad de práctica supervisada y el desempeño satisfactorio en un examen de calificación. Los estatutos para el otorgamiento de la licencia suelen establecer las condiciones disciplinarias para imponer acciones contra los psicólogos, que pueden ir de multas y reprimendas a la suspensión y revocación de la licencia. Muchas jurisdicciones han incorporado en sus estatutos, directa o indirectamente, los requisitos de la APA, por lo que las violaciones al *Código Ético* a menudo constituyen la base para alguna acción disciplinaria. La APA también ha desarrollado una Ley Modelo para el Otorgamiento de la Licencia a los Psicólogos (*Model Act for State Licensure of Psychologists*, APA, 1987b), que sirve como prototipo de las legislaciones estatales que regulan la práctica de la psicología.

A un nivel más avanzado, el Consejo Estadounidense de Psicología Profesional (*American Board of Professional Psychology*, ABPP; véase el apéndice B) otorga una certificación especial. Dado que requiere un nivel mayor de capacitación y experiencia en ciertas especialidades, la ABPP certifica diplomados en áreas como psicología clínica, consejería, industrial/organizacional y educativa por medio de comisiones de especialidades separadas. El Directorio de la APA contiene una lista de los diplomados actuales en cada especialidad, que puede solicitarse también a la ABPP. Como está constituida como una federación privada de comisiones dentro de la profesión, la ABPP no tiene la autoridad legal de la que disponen las dependencias gubernamentales que administran las licencias estatales y las leyes de certificación.

En la última década, los cambios en el sistema estadounidense de salud y otras modificaciones en el mercado profesional dieron lugar a una urgencia creciente en materia de credenciales para la práctica de la psicología. Por consecuencia, la APA ha iniciado acciones para promover el manejo ordenado de los muchos conflictos potenciales inherentes a la atmósfera actual. Una de esas acciones es la creación del Colegio Superior de Psicología Profesional (*College of Professional Psychology*) de la APA (Sleek, 1995), que extenderá las credenciales para diversas habilidades en la psicología mediante un procedimiento que incluye exámenes y requisitos educativos y de experiencia. Otro es el desarrollo de un proceso mediante el cual pueden reconocerse formalmente las especialidades y habilidades en la práctica de la psicología (APA, Comisión Interina Adjunta para la Identificación y Reconocimiento de Especialidades y Habilidades (*Joint Interim Committee for the Identification and Recognition of Specialties and Proficiencies*), 1995a, 1995b). Es indudable que en el futuro próximo veremos una mayor orientación con respecto al tema específico de la capacitación del usuario de las pruebas, así como nuevos procedimientos para acreditar a los especialistas en la evaluación.

RESPONSABILIDADES DE LOS EDITORES DE LAS PRUEBAS

Por lo general, la venta de las pruebas se restringe a quienes cumplen con ciertas condiciones mínimas de capacitación. Los catálogos de las principales editoriales especifican los requisitos que deben satisfacer los compradores, que por lo general consisten en un grado de maestría en psicología o su equivalente. Algunos editores clasifican sus pruebas en niveles, de acuerdo con la capacitación que requieren del usuario, que van de las pruebas de aprovechamiento educativo y destreza vocacional a los tests colectivos de inteligencia, y los inventarios de intereses, los instrumentos clínicos como las pruebas individuales de inteligencia y la mayor parte de los tests de personalidad. También se hacen distinciones entre los compradores particulares y las instituciones autorizadas en la adquisición de las pruebas adecuadas. En el caso de los estudiantes graduados que necesiten determinada prueba para una materia o para una investigación, su profesor debe firmar la orden de compra, con lo que asume la responsabilidad por el uso adecuado del instrumento.[3]

Los esfuerzos por restringir la distribución de las pruebas tienen un objetivo doble: garantizar la seguridad de los materiales del instrumento y prevenir el abuso; sin em-

[3] La Comisión sobre Pruebas Psicológicas y Evaluación (*Committee on Psychological Tests and Assessment*, 1995) de la APA preparó una declaración que orienta sobre el uso seguro de las pruebas psicológicas en la educación de estudiantes graduados y no graduados en psicología.

bargo, debe advertirse que aunque los distribuidores se esfuerzan por cumplir estos objetivos, el control que ejercen es por fuerza limitado. En algunos casos, no pueden investigar y verificar las supuestas capacidades y acreditación de los compradores (véase, por ejemplo, Oles y Davis, 1977), pero, aunque pudieran hacerlo, las capacidades y acreditación formales sólo constituyen un dispositivo general de identificación; por ejemplo, es evidente que un grado de maestría en psicología —o incluso un doctorado, una licencia estatal o un diploma de la ABPP— no necesariamente significa que el individuo está calificado para utilizar una prueba particular o que su formación es relevante para la interpretación apropiada de los resultados obtenidos con el instrumento. La principal responsabilidad por el uso adecuado de las pruebas recae finalmente en el usuario o la institución interesada.

Otra responsabilidad profesional tiene que ver con la venta de instrumentos psicológicos por autores y editoriales. Las pruebas no deben salir prematuramente a la circulación para uso general y no deben hacerse afirmaciones sobre sus posibles méritos si no se cuenta con evidencias objetivas suficientes. Cuando una prueba se distribuye con propósitos exclusivos de investigación, esta condición debe especificarse con claridad y su distribución debe ser restringida. El manual no sólo tiene que incluir datos adecuados que permitan una evaluación del propio instrumento, sino también información completa sobre su aplicación, calificación y normas. Más que una estratagema de ventas diseñada para mejorar la imagen del instrumento, el manual debe ser una exposición fáctica de lo que se sabe sobre la prueba. Es responsabilidad del autor y del editor revisar las pruebas y normas con la frecuencia suficiente para evitar su obsolescencia; desde luego, la rapidez con que una prueba queda rezagada varía con la naturaleza del propio instrumento.

Por razones obvias, no deben publicarse en los medios masivos de comunicación, en parte o del todo, las pruebas cuya seguridad debe garantizarse porque se utilizan para tomar decisiones de selección, colocación o diagnóstico. Cualquier publicidad que se dé a reactivos específicos de la prueba tenderá a invalidar el uso futuro del instrumento con otras personas; además, la publicación de instrumentos en los medios de comunicación puede dar lugar a que la gente haga una autoevaluación que pueda causar daños psicológicos. Otra práctica que casi siempre se considera poco profesional consiste en enviar las pruebas por correo, pues semejante procedimiento no sólo impide controlar las condiciones de la evaluación, sino que en general supone que las puntuaciones se interpreten sin contar con otra información importante del individuo. Salvo muy contadas excepciones, como el uso de inventarios de intereses o valores con individuos medianamente conocedores y motivados, los resultados obtenidos en esas condiciones suelen ser más que inútiles.[4]

Es digno de mención que desde la década de los ochenta las editoriales empezaron a tomar medidas para asegurar que los instrumentos que publican y distribuyen se empleen en forma adecuada y que sus calificaciones se interpreten correctamente. Con este fin han hecho esfuerzos por ampliar y mejorar la comunicación con sus clientes acerca de pruebas específicas y por mejorar la comprensión del público sobre la evaluación en general. Los editores de pruebas psicológicas participaron con la APA y

[4] La Comisión de Ética de la APA preparó recientemente una declaración sobre el hecho de "llevar las pruebas a casa" en respuesta a una investigación concerniente a lo adecuado de enviar a casa el MMPI para su aplicación (APA, Ethics Committee, 1994, pp. 665–666).

otras organizaciones nacionales en los Proyectos de Capacitación y Formación de los Usuarios de las Pruebas de la Comisión Conjunta sobre Prácticas de Evaluación (*Joint Committee on Testing Practices*, Eyde *et al.*, 1988, 1993). Además, establecieron la Asociación de Editores de Tests (*Association of Test Publishers*, ATP; véase el apéndice B), una organización cuyos miembros se comprometieron a promover la integridad de los productos y servicios de evaluación y a mejorar su valor para la sociedad. La ATP publicó recientemente la segunda edición de un conjunto de "Directrices Modelo para la Integridad de la Evaluación Previa a la Contratación" (*Model Guidelines for Preemployment Integrity Testing*; ATP, 1996).

PROTECCIÓN DE LA PRIVACIDAD

Un aspecto particularmente relacionado con las pruebas de personalidad atañe a la invasión de la privacidad. En un informe titulado *Privacy and Behavioral Research* ("Privacidad e Investigación Conductual", 1967) se define el derecho a la privacidad como la prerrogativa para decidir por uno mismo hasta dónde compartir con otros los pensamientos, sentimientos y hechos de la propia vida; este derecho se caracteriza además como "esencial para asegurar la libertad y autodeterminación" (p. 2). En la medida en que algunas pruebas de rasgos emocionales, motivacionales o de actitudes son necesariamente encubiertas, el examinado puede revelar esas características en el curso de la prueba sin percatarse de ello. Aunque puede ser necesario mantener al examinado en la ignorancia de las formas específicas de interpretación de las respuestas a cualquier prueba, a nadie debe someterse a un programa de evaluación con falsas expectativas, por lo que resulta de primordial importancia la obligación de tomar las medidas necesarias para que el examinado comprenda del todo el uso que se hará de los resultados de la prueba.

Aunque la preocupación por la invasión de la privacidad se ha expresado sobre todo con respecto a los tests de personalidad, lógicamente se aplica a cualquier instrumento. Desde luego, cualquier test de inteligencia, de aptitudes o de aprovechamiento puede revelar limitaciones en las habilidades y el conocimiento que el individuo preferiría no revelar; más aún, cualquier observación que se haga de su conducta (como la que puede tener lugar durante una entrevista, una conversación casual o un encuentro personal) puede arrojar información que preferiría ocultar y que revela sin saberlo. El hecho de que a menudo se toque el tema de las pruebas psicológicas en los debates sobre la invasión de la privacidad refleja las ideas engañosas sobre las pruebas así como su frecuente mal uso como la única base para tomar decisiones sobre los individuos. La suspicacia y los temores populares disminuirían si se reconociera que las pruebas sólo son medidas de muestras de conducta que no tienen poderes misteriosos para penetrar más allá del comportamiento. De igual modo, el excesivo énfasis que se pone en el resultado de cualquier instrumento disminuiría si, cuando deben tomarse decisiones importantes para el individuo, las pruebas se interpretaran únicamente en el contexto de evaluaciones comprensivas.

También debería advertirse que cualquier investigación conductual, que utilice pruebas u otros procedimientos de observación, supone la posibilidad de invadir la privacidad. Aun así, como científicos, los psicólogos están comprometidos con la meta de aumentar el conocimiento del comportamiento humano. Es posible que surjan conflictos de valores y que deban ser resueltos en los casos individuales.[5] Como es evi-

dente, no se trata de un problema sencillo, por lo que ha sido objeto de amplias deliberaciones.[6] No pueden formularse reglas universales para salvaguardar la privacidad personal, sólo ofrecerse orientaciones generales. Las soluciones deben plantearse en términos de las circunstancias particulares de cada caso y a la luz de la conciencia ética y la responsabilidad profesional del psicólogo.

Un factor relevante es el propósito por el cual se realiza la evaluación, si se trata de consejería individual, de decisiones institucionales concernientes a la selección y clasificación o de una investigación; por ejemplo, en las situaciones clínicas o de consejería, los clientes suelen estar dispuestos a revelar asuntos personales para obtener ayuda para sus problemas. Cualquiera que sea el propósito de la evaluación, la protección de la privacidad incluye dos conceptos cruciales: la relevancia y el consentimiento informado. La información que se solicita al individuo debe ser *relevante* para los propósitos establecidos de la evaluación, lo que implica que deben hacerse todos los esfuerzos necesarios para asegurar la validez de las pruebas para propósitos predictivos o de diagnóstico particulares para los que se utilizan. Los recientes desarrollos legales, como el caso *Soroka vs. Dayton Hudson* (véase, por ejemplo, Merenda, 1995) y la Ley de 1990 para los Estadounidenses con Discapacidades (P.L. 101–336) subrayaron la importancia de mantener al mínimo la intrusión de las indagatorias en las evaluaciones que se realizan antes de la contratación y de asegurar que sin duda son relevantes para el desempeño laboral (véase, por ejemplo, Bruyère y O'Keeffe, 1994; D. C. Brown, 1996; Herman, 1994, capítulo 2). En el caso *Soroka*, los solicitantes de empleo cuestionaron el uso de una prueba de selección, argumentando que las preguntas relativas a las creencias religiosas y preferencias sexuales (tomadas del MMPI y del CPI) eran intrusivas y discriminatorias. Aunque el caso concluyó sin una resolución definitiva, varios autores de pruebas, incluidos los responsables de las más recientes revisiones del MMPI y del CPI, eliminaron esos reactivos de los inventarios autodescriptivos (véase los capítulos 13 y 17).

El concepto de *consentimiento informado* también requiere ser aclarado, y su aplicación en los casos individuales exige el constante ejercicio del sentido del juicio (AERA, APA, NCME, 1985). No obstante que el *Código Ético* actual contiene una forma explícita exige el consentimiento informado sólo en el caso de la terapia y no para la evaluación, dicho requisito está implícito en otras normas relativas a la evaluación y el diagnóstico en contextos profesionales, así como en muchas otras partes del código. Más aún, las regulaciones de la comisión estatal de psicología, los casos legales, las reglas institucionales o las normas prevalecientes de la práctica requieren por lo común del consentimiento informado en el contexto de las actividades de evaluación y de intervención (Canter *et al.*, 1994, p. 67). Al examinado debe informársele del propósito del examen, de los datos que se buscan y del uso que se hará de las calificaciones; sin embargo, no es necesario que se le muestren de antemano los reactivos de la prueba o que se le diga cómo se calificarán ciertas respuestas. En el caso de los menores,[7] tam-

[5] Los Principios Éticos en el Manejo de la Investigación con Participantes Humanos (*Ethical Principles in the Conduct of Research with Human Participants*, APA, 1982) proporcionan ciertas orientaciones a este respecto.

[6] Véase, por ejemplo, la crítica hecha por F. Allan Hanson (1993) a las pruebas y su función en la sociedad moderna. Aunque el tratado se basa sin duda en la ideología de que la evaluación es antiética y está lejos de ser desapasionada, puede ser de interés para los lectores desde una perspectiva antropológica.

[7] Véase Kamphaus y Frick (1996, capítulo 4) para una orientación sobre el consentimiento para la evaluación y otros aspectos éticos y legales en la evaluación psicológica de los menores.

poco deben mostrarse los reactivos a los padres, ya que dicha información invalidaría la prueba. En el capítulo correspondiente a los derechos de los examinados, los *Estándares de Examinación* consideran todos los aspectos especiales que pueden surgir respecto al consentimiento informado y otros temas relacionados con las situaciones de evaluación.

CONFIDENCIALIDAD

Al igual que la protección de la privaccidad, con la que se relaciona, el problema de la confidencialidad de los datos tiene también muchos aspectos. En respuesta a la interrogante de quién debe tener acceso a los resultados de una prueba pueden influir varias consideraciones, como la seguridad del contenido del instrumento, el riesgo de que se malinterpreten sus resultados y la necesidad de que varias personas los conozcan.

Cada vez se tiene más conciencia del derecho de los individuos a conocer los descubrimientos en los informes de sus tests, a tener la oportunidad de comentar sus contenidos y, de ser necesario, aclarar o corregir la información fáctica. Por su parte, los consejeros intentan que sus clientes participen de manera activa en su propia evaluación. Lo anterior implica la necesidad de presentar los resultados de la prueba de una forma fácilmente comprensible, libre de jerga o etiquetas técnicas y orientada al objetivo inmediato de la evaluación; pero también indica la necesidad de tomar las precauciones adecuadas contra el abuso y la mala interpretación de los resultados del instrumento.

En general, las discusiones sobre la confidencialidad de los registros de las pruebas tratan sobre el acceso de *terceras* personas, distintas del individuo evaluado (o el padre de un menor) y el examinador. El principio del que se parte es que los registros no deberían abrirse sin el conocimiento y el consentimiento del examinado, a menos que se tratara de un mandato legal o que fuera permitido por la ley para un propósito válido. Cuando las pruebas se aplican en un contexto institucional, como el sistema escolar, el tribunal o el medio laboral, en el momento de la evaluación debería informarse al individuo sobre el propósito de la prueba, la forma en que se utilizarán sus resultados y su disponibilidad para el personal de la institución que tenga una necesidad legítima de los mismos. Cuando un extraño solicita los resultados (como sucede cuando un posible empleador o una universidad piden los resultados de la prueba al sistema escolar), se requiere un consentimiento expreso para entregar los datos. Los mismos requisitos se aplican a las pruebas aplicadas en los contextos clínicos y de consejería o cuando se realizan con propósitos de investigación. En la Declaración sobre la revelación de datos de pruebas (*Statement on the Disclosure of Test Data*, APA, 1996) preparada por la CPTA para ayudar a los psicólogos a manejar las cuestiones concernientes a la revelación de los registros de las pruebas, puede encontrarse mayor orientación al respecto. Además, la Comisión sobre Asuntos Legales de la APA (*Committee on Legal Issues*, COLI) ha establecido ciertas estrategias para psicólogos que deben enfrentar emplazamientos o que en el curso de su práctica profesional se vean obligados a prestar testimonio concerniente a los registros o los datos de la evaluación psicológica de algún cliente (APA, COLI, 1996).

Otro problema tiene que ver con la conservación de los registros en las instituciones. Por un lado, los registros longitudinales de los individuos pueden resultar muy valiosos, no sólo para propósitos de investigación, sino también para comprenderlos y aconsejarlos, lo que desde luego presupone el uso y la interpretación adecuados de los resultados de la prueba. Pero, por otro lado, la disponibilidad de registros antiguos hace posible que se cometan abusos como las inferencias incorrectas de datos obsoletos y el acceso no autorizado para propósitos diferentes a los de la evaluación original; por ejemplo, sería absurdo citar el CI o la puntuación obtenida por un niño de tercer grado en una prueba de lectura cuando se lo evalúa para su admisión a la universidad. De modo similar, cuando los registros se conservan durante muchos años, existe el riesgo de que puedan utilizarse para propósitos que el examinado (o sus padres) nunca anticiparon y que podrían no haber aprobado. Para evitar tales abusos, cuando los registros se conservan ya sea para un uso longitudinal legítimo del interés del individuo o por propósitos aceptables de investigación, el acceso a los mismos debe estar sujeto a controles estrictos. Cada institución debe formular políticas explícitas respecto a la destrucción, conservación y acceso a los registros personales. La declaración al respecto de las Directrices para el mantenimiento de los registros (*Record Keeping Guidelines*; APA, COPPS, 1993), contiene más información sobre este tema.

COMUNICACIÓN DE LOS RESULTADOS DE LAS PRUEBAS

En los últimos años los psicólogos empezaron a prestar más atención a la necesidad de comunicar los resultados de las pruebas de tal forma que resultaran significativos y útiles para el sujeto. Es claro que la información no debe transmitirse de una manera rutinaria ni proporcionar las explicaciones interpretativas apropiadas. Hay que preferir las descripciones cualitativas, hechas en términos sencillos de amplios niveles de desempeño, a las calificaciones numéricas específicas, excepto cuando la comunicación es con profesionales adecuadamente capacitados. Se sabe que incluso la gente con un buen nivel de educación confunde percentiles con porcentajes, percentiles con CI, normas con estándares y clasificaciones de intereses con calificaciones de aptitudes. Pero una interpretación errónea más grave tiene que ver con las conclusiones extraídas de las puntuaciones de una prueba, incluso cuando el significado técnico se entienda correctamente. Un ejemplo es la suposición popular de que el CI indica una característica fija del individuo que predetermina su nivel permanente de aprovechamiento intelectual.

Entre los posibles receptores de los resultados de las pruebas, además de los propios examinados, están los padres de los menores, maestros y otros miembros del personal escolar, empleadores, psiquiatras y personal de tribunales y correccionales. En todas las comunicaciones relacionadas con la evaluación, es deseable tener en cuenta las características de quien recibe la información, lo que no sólo concierne a su nivel general de educación y sus conocimientos de la psicología y de la evaluación, sino también a la respuesta emocional anticipada a la información; en el caso de un padre o un maestro, digamos, la relación emocional que tienen con el niño puede interferir con la aceptación serena y racional de la información fáctica.

Al comunicar los resultados de las pruebas a los propios examinados, se enfrentan problemas similares, trátese de niños o de adultos.[8] En este caso, se aplican las mismas precauciones generales contra las malas interpretaciones que en la comunicación a terceras partes. A este respecto, los *Estándares de Examinación* insisten en la necesidad de que quienes usan las pruebas en aplicaciones clínicas y de consejería den a los examinados explicaciones apropiadas y comprensibles de los resultados y cualquier recomendación que pueda surgir de ellas.

La consideración de las reacciones emocionales a la información sobre las pruebas es especialmente importante cuando los individuos se enteran de sus propias ventajas y desventajas. Al entregarles los resultados de sus pruebas, deben tomarse las medidas que aseguren que los datos fueron interpretados en forma apropiada por una persona calificada, pero también es importante proporcionar consejería a cualquiera que pueda resultar emocionalmente perturbado por la información; por ejemplo, un estudiante universitario puede quedar muy desanimado cuando se entera de su mal desempeño en una prueba de aptitud escolar; un estudiante sobredotado puede desarrollar hábitos de descuido y pereza o volverse difícil y poco cooperativo si descubre que es mucho más brillante que sus compañeros. Incluso cuando la prueba ha sido aplicada, calificada e interpretada de manera adecuada, el conocimiento de una calificación sin tener la oportunidad de hablar de él puede resultar nocivo para el individuo.

Los psicólogos consejeros se han preocupado por desarrollar formas eficaces de transmitir la información de las pruebas a sus clientes (véase, por ejemplo, Hood y Johnson, 1997, capítulo 17), y aunque los detalles del proceso escapan a nuestra revisión, resultan de particular interés dos líneas principales. Primero, el informe del instrumento debe considerarse como parte integral del proceso de consejería e incorporarse a la relación entre cliente y consejero. Segundo, hasta donde sea posible, es conveniente que los consejeros comprometan a sus clientes en la interpretación de los resultados de las pruebas en función de las preguntas específicas que puedan hacer. Una consideración importante en la consejería se relaciona con el hecho de que el cliente acepte la información que se le presenta. La situación de consejería es tal, que si por cualquier razón el individuo rechaza alguna información es probable que ésta sea totalmente desechada. Por otro lado, la aceptación de los datos bien interpretados de la prueba puede ser por sí misma de valor terapéutico para el cliente, en especial en el contexto del tratamiento con orientación cognoscitiva.

EVALUACIÓN EN DIVERSAS POBLACIONES

El entorno. Las décadas transcurridas desde los cincuenta han presenciado un aumento en la preocupación del público por los derechos de las minorías étnicas, las mujeres y los individuos con discapacidades, además de otros grupos minoritarios.[9] Esta

[8] El número de septiembre de 1992 de *Psychological Assessment* contiene una sección especial sobre el tema de proporcionar a los clientes realimentación sobre las pruebas psicológicas. En un artículo particularmente útil, K. S. Pope analiza diez aspectos fundamentales de la realimentación que, según afirma, "puede ser el aspecto más descuidado de la evaluación" (1992, p. 265).

[9] Aunque las mujeres representan una mayoría estadística en la población de los Estados Unidos, legal y ocupacionalmente, entre otras formas, han compartido muchos de los problemas de las minorías; de ahí que al usar el término "minoría" se entiende la inclusión de las mujeres.

preocupación se refleja en la promulgación de la legislación de los derechos civiles, tanto a nivel federal como estatal. Las pruebas psicológicas han sido un foco especial de atención en lo que respecta a los mecanismos para mejorar las oportunidades educativas y vocacionales de los individuos de esos diversos grupos (Gifford, 1989a, 1989b). La bibliografía psicológica contiene muchos análisis del tema cuyo impacto fluctúa entre la aclaración y la ofuscación; entre los primeros se encuentran diversos trabajos presentados por las asociaciones profesionales (véase, por ejemplo, ACA, 1989; APA, Board of Ethnic Minority Affairs, 1990; APA, Division of Evaluation, Measurement, and Statistics, 1993; Prediger, 1993; Sackett y Wilk, 1994). Además, cada vez es mayor el acceso a la orientación sobre las prácticas adecuadas de evaluación con poblaciones diversas (véase el capítulo 9; Dana, 1996a; Sattler, 1988, capítulos 19 y 20; Suzuki *et al.*, 1996; Valencia y López, 1992). Por otro lado, los informes preparados bajo los auspicios del Consejo Nacional Estadounidense para la Investigación (*National Research Council*), la Oficina de Evaluación Tecnológica (*Office of Technology Assessment*) y otros grupos citados en una sección anterior del capítulo han examinado las controversias sobre las pruebas a la luz del contexto social actual y han presentado opiniones equilibradas de las funciones de la evaluación.

En buena medida, la preocupación tiene que ver con la disminución de las calificaciones en las pruebas que obedecen a condiciones culturales que podrían haber afectado el desarrollo de aptitudes, intereses, motivación, actitudes y otras características psicológicas de los miembros de los grupos minoritarios. Algunas de las soluciones propuestas al problema reflejan una concepción errónea de la naturaleza y la función de las pruebas psicológicas. Las diferencias en los antecedentes de la experiencia de grupos o individuos se manifiestan inevitablemente en la ejecución de las pruebas. Cada instrumento psicológico mide una muestra de conducta, y la influencia de la cultura debe ser detectada por las pruebas en la medida en que afecte al comportamiento. Si pudiéramos descartar todos los diferenciales culturales de las pruebas, disminuiríamos su validez como medida del área conductual que pretende evaluar, en cuyo caso, la herramienta no lograría proporcionar la información necesaria para corregir las mismas condiciones que dañaron la ejecución.

En los capítulos 9 y 12 revisamos con mayor amplitud tanto los procedimientos como la lógica teórica que fundamentan la evaluación de diferentes clases de poblaciones especiales; en el capítulo 6 hicimos un análisis técnico del concepto de "sesgo de la prueba" en conexión con la validez del instrumento. En este capítulo, nuestro interés se concentra principalmente en los aspectos profesionales y las implicaciones sociales de la evaluación de grupos minoritarios.

Regulaciones legales. Desde 1960, se observa en los Estados Unidos, una serie de desarrollos relacionados con la evaluación educativa y laboral de las minorías que comprende acciones legislativas, órdenes ejecutivas y decisiones legales. Explicamos las legislaciones correspondientes a la evaluación educativa en los capítulos 9 y 17, y el lector interesado puede encontrar en Linn y Gronlund (1995, capítulo 18)[10] una revisión de los temas y tendencias actuales en la evaluación decretada externamente.

[10] El lector puede encontrar en Ayers, Day y Rotatori (1990) y en Reschly (1988) las revisiones de algunas decisiones legales importantes en el campo de la evaluación psicoeducativa.

El papel que los tribunales desempeñan en el campo laboral es cada vez más destacado en la interpretación y aplicación de las leyes sobre los derechos civiles, cuyas implicaciones han sido analizadas en la bibliografía correspondiente por personas con formación en psicología, leyes o ambas (véase, por ejemplo, APA, CPTA, 1988; Bersoff, 1983, 1984; Bruyère y O'Keeffe, 1994; Hollander, 1982; Merenda, 1995; Meyers, 1992; Wigdor, 1982). La legislación federal más relevante la proporciona el título VII de la Ley de Derechos Civiles de 1964 (P.L. 88–352), conocida también como la Ley de Igualdad de Oportunidades en el Empleo (*Equal Employment Opportunity Act*) con sus enmiendas subsecuentes: la Ley de Derechos Civiles de 1991 (P.L. 102–166) y la Ley de Estadounidenses con Discapacidades de 1990 (P.L. 101–336). La responsabilidad por la implantación y el cumplimiento de estas leyes corresponde principalmente a la EEOC, que con ese propósito establece y distribuye directrices. En aras de simplificar el procedimiento y mejorar la coordinación entre instituciones, la EEOC, la Comisión de Servicio Civil (ahora conocida como La Oficina de Manejo de Personal de los Estados Unidos, *Office of Personnel Management*) y los departamentos de Justicia, de Trabajo y del Tesoro[11] adoptaron en 1978 las Directrices Uniformes para los Procedimientos de Selección de Empleados (*Uniform Guidelines on Employee Selection Procedures*).

La Ley de Igualdad de Oportunidades para el Empleo prohíbe la discriminación por raza, color, religión, sexo y origen nacional en los procedimientos de selección que conducen a las decisiones de empleo. Estas regulaciones se aplican a los empleadores (privados o gubernamentales), organizaciones laborales, oficinas de empleo y comisiones para el otorgamiento de licencias y certificación. Cuando el uso de una prueba u otro procedimiento de selección da por resultado una tasa de rechazo sustancialmente más alta para los candidatos de la minoría que para los candidatos de grupos no minoritarios ("impacto adverso"),[12] es necesario justificar su utilidad mediante evidencias de su validez para el trabajo en cuestión.

Históricamente, los requisitos para la validación de pruebas aceptables fueron establecidos mediante los *Estándares de Examinación*, los "Principios para la validación y uso de los procedimientos de selección de personal" (*Principles for the Validation and Use of Personnel Selection Procedures*, SIOP, 1987) y algunos otros documentos de la profesión; sin embargo, en las dos últimas décadas tuvieron lugar varios casos en los que consideraciones legales extrañas irrumpieron en las prácticas psicométricas, especialmente en relación con los derechos civiles. Uno de estos casos se conoce como la "Regla de Oro" (véase también el capítulo 7) y surgió de una disputa entre la compañía de seguros Golden Rule y el Servicio de Evaluación Educativa (*Educational Testing Service*, ETS) concerniente a los exámenes preparados por el ETS para otorgar licencias a los agentes de seguros. El arreglo establecía que cuando se encontraran tasas diferenciales de respuestas correctas entre los grupos minoritarios y mayoritarios, debería darse prioridad al uso de reactivos que dieran lugar a las menores diferencias entre grupos; aunque la medida pretendía promover la equidad y disminuir al mínimo el

[11] Las Directrices Uniformes (*Uniform Guidelines*) han quedado rezagadas y es evidente que necesitan revisarse. Se espera contar con una nueva versión de las mismas después de la publicación de los nuevos *Estándares de Examinación*, cuya publicación se anticipa para finales de la década de los noventa (véase el capítulo 1).

[12] B. Lerner (1980a) analizó las inconsistencias en la forma de evaluar el impacto adverso en diferentes casos legales (véase también Ironson, Guion y Ostrander, 1982).

impacto adverso, el acuerdo de la Regla de Oro motivó un acalorado debate en torno a sus suposiciones sobre la naturaleza del sesgo del reactivo y al grado al que la evidencia empírica justifica el procedimiento sugerido por el acuerdo (APA, CPTA, 1988; Bond, 1987; Linn y Drasgow, 1987; Rooney, 1987).

En la revisión de la acción afirmativa, las *Directrices uniformes* adoptadas en 1978 señalan que incluso cuando los procedimientos de selección han sido validados satisfactoriamente, si se obtienen tasas de rechazo desproporcionadas para las minorías, deben tomarse medidas para reducir esta discrepancia en la medida de lo posible. *Acción afirmativa* implica que las empresas deben hacer algo más que evitar las prácticas discriminatorias. En términos psicológicos, puede considerarse que los programas de acción afirmativa, que en los últimos años han sido objeto de ataques crecientes en los medios políticos, son el esfuerzo por compensar los efectos residuales de las anteriores inequidades sociales. La costumbre de establecer normas para subgrupos, puesta en práctica con el GATB en la década de los ochenta para producir tasas comparables de empleo para solicitantes blancos, negros e hispanos a pesar de las puntuaciones obtenidas en las pruebas de aptitud (capítulo 17; véase también Hartigan y Wigdor, 1989), constituyó un ejemplo de la acción afirmativa destinada a reducir el impacto adverso que las pruebas podrían tener en la contratación; sin embargo, generó tal controversia que llevó a la aprobación de la Ley de Derechos Civiles de 1991 (P.L. 102–166) que prohíbe explícitamente cualquier forma de ajuste a las puntuaciones en función de raza, color, religión, sexo u origen nacional. En el campo de la evaluación psicológica, se ha reconocido que las ramificaciones de la mencionada ley "tienen un alcance mucho mayor que el previsto por el Congreso" (D. C. Brown, 1994, p. 927) y podría restringir gravemente el uso de tests de personalidad y de pruebas de capacidad física que utilizan normas distintas para hombres y mujeres (véase también L. S. Gottfredson, 1994; Kehoe y Tenopyr, 1994; Sackett y Wilk, 1994). De hecho, algunos autores y editores de pruebas ya han tomado medidas para ofrecer procedimientos de calificación alternativos que eliminen la separación de normas por género (véase, por ejemplo, Gough y Bradley, 1996).

La Ley de los Estadounidenses con Discapacidades (*American with Disabilities Act*, ADA) de 1990 (P.L. 101–336) representa otro esfuerzo bien intencionado por eliminar las barreras a la equidad de oportunidades que ha generado preocupación en empleadores y otras personas interesadas en las prácticas adecuadas de evaluación para la contratación. Las cláusulas de empleo de la ADA impiden que los empleadores utilicen pruebas médicas o indaguen acerca de los antecedentes de abuso de drogas o las condiciones psiquiátricas antes de hacer una oferta de trabajo. Las directrices y regulaciones de la EEOC sobre las indagaciones previas a la contratación relacionadas con la discapacidad y los exámenes médicos (1994, 1995) han dejado sin resolver la cuestión de qué instrumentos de medición psicológicos y de personalidad son admisibles en las situaciones relacionadas con la contratación.

Es probable que en el futuro se encuentren faltas de correspondencia entre los mandatos éticos, legales y profesionales (véase, por ejemplo, D. C Brown, 1996) que indudablemente harán más difícil la aplicación de pruebas psicológicas a la toma de decisiones en las llamadas áreas de "alto riesgo" del empleo y la educación. En gran medida, esta situación representa un avance porque destaca la necesidad de reconocer explícitamente la participación de los valores en la toma de cualquier decisión que tenga alguna consecuencia, sea en el campo científico o en la práctica. En palabras de

Messick, "los valores son intrínsecos al significado y a los resultados de la evaluación, y siempre lo han sido [...] Esto hace explícito lo que siempre ha estado latente, a saber: que los juicios sobre la validez son juicios de valor" (1995, p. 748). Con todo, incluso la gente bien intencionada y razonable puede y, de hecho, lo hace, tener grandes discrepancias acerca de los valores, y es ahí donde está la dificultad.

Factores relacionados con las pruebas. Al examinar a diversas personas es importante diferenciar entre los factores que influyen en la prueba y el criterio conductual y aquellos cuyo efecto se restringe al instrumento. Son los relacionados con la prueba los que reducen la validez. Los ejemplos incluyen la experiencia en la presentación de pruebas, la motivación para tener un buen desempeño, el *rapport* establecido con el examinador, el énfasis indebido en la rapidez y cualquier otra variable que afecte la ejecución en pruebas particulares pero que sean irrelevantes para el área general de conducta considerada. Al evaluar a individuos con antecedentes culturales disímiles o con discapacidades deben hacerse esfuerzos especiales para reducir la operación de los factores relacionados con las pruebas (véase, Sattler, 1988, capítulos 19 y 20). A este respecto, resulta sumamente deseable contar con procedimientos adecuados de orientación para la presentación del instrumento, con práctica preliminar y otros pasos desarrollados con ese propósito (capítulos 1 y 9).

El contenido del instrumento también puede influir en las puntuaciones de formas no relacionadas con la capacidad que la prueba pretende medir; por ejemplo, en una prueba de razonamiento aritmético el uso de nombres o imágenes de objetos poco familiares para cierto medio cultural representan una desventaja restringida al instrumento. Otra forma más sutil en que el contenido del instrumento puede afectar espuriamente el desempeño es mediante las actitudes y respuestas emocionales del examinado; por ejemplo, las historias o imágenes que presentan escenas de una clásica familia de clase media que vive en los suburbios puede perturbar a un niño criado en una barriada en un hogar de bajos ingresos. La perpetuación de estereotipos sexuales en el contenido de las pruebas, como presentar a los hombres como doctores o pilotos y a las mujeres como enfermeras o aeromozas, también puede tener efectos nocivos. A la luz de estas consideraciones, casi todas las editoriales hacen ahora esfuerzos especiales por eliminar los contenidos inadecuados de sus instrumentos. De hecho, en el proceso de elaboración del instrumento es común la revisión del contenido en atención a las posibles implicaciones negativas para los examinados de grupos minoritarios (véase, por ejemplo, *ETS Standards*, 1981/1987).

La evaluación de personas con diversos antecedentes culturales e historial de experiencias, así como de los individuos con discapacidades, es una preocupación evidente de los *Estándares de Examinación*; esta orientación general se refleja en diversos criterios individuales para el desarrollo y uso de las pruebas. Además, capítulos especiales, con sus propios conjuntos de estándares, manejan los problemas de la evaluación de personas con condiciones discapacitantes y diferencias lingüísticas, que representan segmentos considerables de la población de los Estados Unidos.

Interpretación y uso de las puntuaciones de las pruebas. Con mucho, la consideración más importante en la evaluación de grupos diversos (como en la evaluación en general) atañe a la interpretación de las puntuaciones de las pruebas. Los

recelos más frecuentes relacionados con la aplicación de estos instrumentos a los grupos minoritarios surgen de la mala interpretación de las puntuaciones. Si un miembro de un grupo minoritario obtiene una baja puntuación en un test de aptitud o una puntuación desviada en uno de personalidad, es esencial investigar a qué se debe. Una puntuación baja en una prueba de aritmética podría, digamos, ser resultado, entre otras razones, de una pobre motivación, de poca habilidad en la lectura o del conocimiento inadecuado de la aritmética. Pero también debería considerarse el tipo de normas que deben emplearse al evaluar las puntuaciones individuales.[13]

Las pruebas están diseñadas para demostrar lo que un individuo puede hacer en un momento determinado, pero no nos pueden decir *por qué* se desempeña como lo hace. Debemos buscar la respuesta a esa pregunta en sus antecedentes, motivaciones y otras circunstancias relevantes. Las pruebas tampoco pueden decirnos lo capaz que podría haber sido un niño criado con desventajas culturales o educacionales si hubiera contado con un ambiente más favorable; tampoco pueden compensar la privación cultural, eliminando sus efectos de las puntuaciones. Por el contrario, las pruebas deberían revelar esos efectos de forma que pudieran tomarse las medidas de remedio convenientes. Tratar de ocultar los efectos de las desventajas culturales, rechazando las pruebas o tratando de elaborar otras que no sean susceptibles a esos efectos, sólo puede retardar el progreso hacia una genuina solución de los problemas sociales.

Todavía es muy común la tendencia a categorizar y etiquetar, como sustituto de la comprensión. Un ejemplo bien conocido se encuentra en las categorías de diagnóstico de la psiquiatría clásica, en la que se etiquetaba a los pacientes como "esquizofrénico paranoide" o "maniaco-depresivo". Conscientes de las muchas deficiencias de ese sistema de clasificación, los autores de los más recientes manuales de diagnóstico psiquiátrico describen trastornos de varios tipos y asignan las etiquetas a las condiciones más que a los individuos que los padecen (véase, por ejemplo, American Psychiatric Association, 1994). Más aún, cada vez son más los psicólogos que hacen descripciones de personalidad que, a diferencia de las etiquetas de diagnóstico, se concentran en los orígenes y el significado individual de la conducta desviada y proporcionan una base más efectiva para la terapia; sin embargo, no resulta sencillo deshacerse de las etiquetas tradicionales.

Otro ejemplo de la tendencia a la categorización se encuentra en las interpretaciones erróneas del CI, que según las creencias populares, es un indicador del potencial intelectual innato y representa una propiedad fija del organismo. Como vimos en el capítulo 12, esta visión no es teóricamente defendible ni la respaldan datos empíricos. Cuando se interpretan adecuadamente, las puntuaciones de los tests de inteligencia no deben fomentar una categorización rígida de las personas, por el contrario, este tipo de pruebas (como cualquier otro instrumento) pueden considerarse como un mapa en el que es posible localizar la posición actual del individuo. Cuando se combinan con la información sobre los antecedentes de su experiencia, las puntuaciones de las pruebas deberían facilitar la planeación adecuada para el desarrollo óptimo del individuo.

Objetividad de las pruebas. Cuando los estereotipos sociales y los prejuicios pueden distorsionar las evaluaciones interpersonales, las pruebas proporcionan un amparo

[13] Una sección especial del número de diciembre de 1994 de *Psychological Assessment* se dedicó a proporcionar información y orientación sobre diversos aspectos de la evaluación normativa.

contra el favoritismo y las decisiones arbitrarias y caprichosas. A medida que los movimientos por los derechos civiles ganaron impulso, diversos observadores llamaron la atención sobre la función positiva que podía cumplir la evaluación estandarizada. Al comentar sobre el uso de las pruebas en las escuelas, J. W. Gardner (1961, pp. 48–49) escribía: "Las pruebas no pueden ver si el joven vestía harapos o paño, ni pueden escuchar el acento de los barrios bajos. Las pruebas revelan las dotes intelectuales en cada nivel de la población".

Aunque las pruebas fueran eliminadas permanecería la necesidad de que individuos e instituciones tomen decisiones, por lo cual tendrían que retroceder a las conocidas alternativas de las cartas de recomendación, entrevistas y promedios de calificación. En la actualidad, es común que estas otras fuentes de información se utilicen junto con las puntuaciones de las pruebas, pero no en su lugar. De hecho, las pruebas estandarizadas se introdujeron como una forma de compensar la poca confiabilidad, la subjetividad y los posibles sesgos de los procedimientos tradicionales. Las alternativas de evaluación generalmente han demostrado ser menos exactas que las pruebas para predecir el desempeño en la escuela o el trabajo (Wigdor y Garner, 1982; pto. I, capítulo 1). Otros procedimientos desarrollados más recientemente, como las técnicas de evaluación del desempeño y los llamados portafolios, quizá demuestren ciertas ventajas cuando se los compare con los instrumentos tradicionales; sin embargo, hasta ahora la investigación realizada con estas técnicas indica que no son más válidas ni más imparciales con las diversas poblaciones que las pruebas estandarizadas a las que podrían complementar o reemplazar (véase el capítulo 17).

Con frecuencia, los ataques a las pruebas no logran diferenciar entre las contribuciones positivas de la evaluación y la imparcialidad en la toma de decisiones y los usos inadecuados de tales instrumentos como sustitutos del juicio cuidadosamente considerado. Al examinar la evaluación en su contexto social, la Comisión para la Evaluación de la Habilidad (*Committee on Ability Testing;* Wigdor y Garner, 1982, pto. I) advertía que las pruebas no deberían considerarse como panaceas ni como chivos expiatorios para los problemas de la sociedad, y que las metas sociales de aumentar las oportunidades para los miembros de los grupos minoritarios no deberían confundirse con la validez del proceso de evaluación. En una de las conclusiones, la comisión observó: "La búsqueda de una sociedad más equitativa ha colocado a la evaluación de la habilidad en el centro de la controversia y le ha dado una reputación exagerada para bien y para mal" (p. 239). Esta declaración sigue siendo válida y, a la luz de la escasez de alternativas viables, es de creer que seguirá siéndolo durante mucho tiempo.

En suma, es indudable que las pruebas pueden utilizarse mal con las minorías, como con cualquiera; empero, cuando se utilizan adecuadamente, cumplen una función importante al impedir la discriminación irrelevante e injusta. Cuando se evalúan las consecuencias sociales de la examinación es necesario considerar con cuidado las consecuencias sociales de *no* realizarla y de tener que confiar en otros procedimientos menos equitativos que la evaluación para tomar decisiones; más aún, al determinar sus consecuencias es preciso diferenciar las repercusiones del uso adecuado de las pruebas de las que pueden resultar de una inadecuada utilización y separar las consecuencias directas de la evaluación de las que son mediadas por factores ajenos al instrumento (Tenopyr, 1995). De otro modo, es probable que descartemos, por razones equivocadas, un instrumento que, aunque siempre es perfectible, puede resultar irremplazable.

Apéndice A

Lista por orden alfabético de pruebas y otros instrumentos de evaluación

En esta tabla presentamos la lista de todos los instrumentos revisados o citados en el texto, a excepción de: (a) los que ya salieron de circulación y que solamente mencionamos por razones históricas, (b) los que ya no se publican, y (c) los descritos en la bibliografía y a los que sólo se tiene acceso recurriendo a sus autores. Para mayor información, consulte los Anuarios de Medición Mental (*Mental Measurements Yearbooks*, MMY) y otras publicaciones del Instituto Buros de Mediciones Mentales, *Test Critiques* (Keyser y Sweetland, 1984–1994), lo mismo que otras fuentes mencionadas en el capítulo uno.

NOMBRE DE LA PRUEBA (ABREVIATURA)/CÓDIGO DEL EDITOR[1]

AAMR Adaptive Behavior Scale (ABS) / PRO-ED
ACT Assessment / ACT
Adjective Check List (ACL) / CPP
Alcohol Use Inventory (AUI) / NCS
Armed Forces Qualification Test (AFQT) / U.S. Military
Armed Services Vocational Aptitude Battery (ASVAB) / U.S. Military
Basic Personality Inventory (BPI) / Sigma
Bayley Infant Neurodevelopmental Screener (BINS) / TPC
Bayley Scales of Infant Development—Second Edition (Bayley-II) / TPC
Beck Depression Inventory (BDI) / TPC
Behavior Assessment System for Children (BASC) / AGS
Bender Visual Motor Gestalt Test (Bender-Gestalt) / WPS

[1] El apéndice B contiene los nombres y las direcciones de las editoriales, dispuestos en orden alfabético de acuerdo con los códigos anotados aquí.

Bennett Mechanical Comprehension Test (BMCT) / TPC
Benton Visual Retention Test, Fifth Edition (BVRT) / TPC
Boehm Test of Basic Concepts—Revised (Boehm-R) / TPC
Bracken Basic Concept Scale (BBCS) / TPC
Brief Symptom Inventory (BSI) / NCS
British Ability Scales (BAS) / NFER-Nelson
Bruininks-Oseretsky Test of Motor Proficiency / AGS
California Achievement Tests—Fifth Edition (CAT) / CTB
California Child Q-Set / CPP
California Diagnostic Mathematics Tests (CDMT) / CTB
California Diagnostic Reading Tests (CDRT) / CTB
California Psychological Inventory—Third Edition (CPI-3) / CPP
California Q-Sort Deck / CPP
Campbell Interest and Skill Survey (CISS) / NCS
Career Assessment Inventory—The Enhanced Version (CAI-EV) / NCS
Career Assessment Inventory—The Vocational Version (CAI-VV) / NCS
Career Beliefs Inventory (CBI) / CPP
Career Development Inventory (CDI) / CPP
Career Directions Inventory (CDI) / Sigma
Career Interest Inventory (CII) / TPC
Career Planning Program (CPP) / ACT
Children's Apperception Test (C.A.T.) / CPS
Cognitive Abilities Test (CogAT, Form 5) / Riverside
College Level Examination Program (CLEP) / ETS
Columbia Mental Maturity Scale (CMMS) / TPC
Comprehensive Tests of Basic Skills—Fourth Edition (CTBS/4) / CTB
Computer Programmer Aptitude Battery (CPAB) / SRA
Concept Assessment Kit—Conservation (CAK) / EdITS
Crawford Small Parts Dexterity Test (CSPDT) / TPC
Cross-Cultural Adaptability Inventory (CCAI) / NCS
CRT Skills Test / SRA
Culture Fair Intelligence Test/IPAT
Das-Naglieri Cognitive Assessment System (CAS) / Riverside
Differential Ability Scales (DAS) / TPC
Differential Aptitude Tests—Computerized Adaptive Edition (DAT Adaptive) / TPC
Differential Aptitude Tests—Fifth Edition (DAT) / TPC
Draw-a-Man Test (see Goodenough-Harris Drawing Test)
Edwards Personal Preference Schedule (EPPS) / TPC
Embedded Figures Test (EFT) / CPP
Fagan Test of Infant Intelligence / Infantest
Fleishman Job Analysis Survey (F-JAS) / MRI
General Aptitude Test Battery (GATB) / USES
Goodenough-Harris Drawing Test / TPC
Graduate Record Examinations (GRE) / ETS
Group Embedded Figures Test / CPP
Guide to the Assessment of Test Session Behavior for the WISC-III and the WIAT / TPC
Guilford-Zimmerman Temperament Survey (GZTS) / CPP
Halstead-Reitan Neuropsychological Test Battery (HRB) / RNL
Haptic Intelligence Scale / Stoelting

Harrington-O'Shea Career Decision-Making System—Revised (CDM-R) / AGS
Hogan Personality Inventory—Second Edition (HPI) / HAS
House-Tree-Person (H-T-P) / WPS
Infant-Toddler Developmental Assessment (IDA) / Riverside
Iowa Tests of Basic Skills / Riverside
Iowa Tests of Educational Development / Riverside
Jackson Personality Inventory—Revised (JPI-R) / Sigma
Jackson Vocational Interest Survey (JVIS) / Sigma
Kaufman Adolescent and Adult Intelligence Test (KAIT) / AGS
Kaufman Assessment Battery for Children (K-ABC) / AGS
Kaufman Brief Intelligence Test (K-BIT) / AGS
Kaufman Test of Educational Achievement (K-TEA) / AGS
Kuder General Interest Survey (KGIS) / CTB
Kuder Occupational Interest Survey (KOIS) / CTB
Kuder Preference Record—Vocational (KPR-V) / CTB
Leiter International Performance Scale—Revised (LIPS-R) / Stoelting
Luria-Nebraska Neuropsychological Battery (LNNB) / WPS
Machover Draw-a-Person Test (D-A-P) / Thomas
McCarthy Scales of Children's Abilities (MSCA) / TPC
Metropolitan Achievement Test—Seventh Edition (MAT) / TPC
Metropolitan Readiness Tests—Sixth Edition (MRT) / TPC
MicroCog: Assessment of Cognitive Functioning / TPC
Millon Adolescent Clinical Inventory (MACI) / NCS
Millon Adolescent Personality Inventory (MAPI) / NCS
Millon Clinical Multiaxial Inventory-III (MCMI-III) / NCS
Millon Index of Personality Styles (MIPS) / TPC
Minnesota Clerical Test (MCT) / TPC
Minnesota Multiphasic Personality Inventory—Adolescent (MMPI-A) / UMP
Minnesota Multiphasic Personality Inventory—2 (MMPI-2) / UMP
Multidimensional Aptitude Battery (MAB) / Sigma
Myers-Briggs Type Indicator (MBTI) / CPP
Occupational Aptitude Survey and Interest Schedule—Second Ed. (OASIS-2) / PRO-ED
Ordinal Scales of Psychological Development / UIP
Orleans-Hanna Algebra Prognosis Test / TPC
Otis-Lennon School Ability Test—Seventh Edition (OLSAT7) / TPC
Paced Auditory Serial Addition Test (PASAT) / ForThought
PDI Customer Service Inventory / PDI
PDI Employment Inventory / PDI
Peabody Picture Vocabulary Test—Revised (PPVT-R) / AGS
Personality Assessment Inventory (PAI) / PAR
Personality Inventory for Children—Revised (PIC-R) / WPS
Personality Inventory for Youth (PIY) / WPS
Personality Research Form (PRF) / Sigma
Porteus Mazes / TPC
Quality of Life Inventory (QOLI) / NCS
Raven's Progressive Matrices (RPM) / Oxford (U.S. Distributor: TPC)
Revised Minnesota Paper Form Board Test (RMPFBT) / TPC
Revised NEO Personality Inventory (NEO-PI-R) / PAR
Roberts Apperception Test for Children (RATC) / WPS

Rorschach / H & H
Rotter Incomplete Sentences Blank (ISB) / TPC
Schaie-Thurstone Adult Mental Abilities Test / CPP
Schedule for Nonadaptive and Adaptive Personality (SNAP) / UMP
Scholastic Assessment Test (SAT) / ETS
Self-Directed Search (SDS) / PAR
Senior Apperception Test (S.A.T.) / CPS
Sixteen Personality Factor Questionnaire—Fifth Edition (16PF) / IPAT
Social Climate Scales / CPP
Social Skills Rating System (SSRS) / AGS
Stanford Achievement Test—Eighth Edition / TPC
Stanford-Binet Intelligence Scale—Fourth Edition (SB-IV) / Riverside
Stanford Diagnostic Mathematics Test—Third Edition (SDMT) / TPC
Stanford Diagnostic Reading Test—Third Edition (SDRT) / TPC
Stanford Writing Assessment Program / TPC
State-Trait Anger Expression Inventory (STAXI) / PAR
State-Trait Anxiety Inventory (STAI) / CPP
State-Trait Anxiety Inventory for Children (STAIC) / CPP
Strong Interest Inventory (SII) / CPP
Structure of Intellect Learning Abilities Test / WPS
Student Adaptation to College Questionnaire (SACQ) / WPS
Student Self-Concept Scale (SSCS) / AGS
Student Styles Questionnaire / TPC
Symptom Checklist-90—Revised (SCL-90-R) / NCS
System for Interactive Guidance Information—Revised (SIGI-PLUS) / ETS
Teamwork-KSA / SRA
Tell-Me-A-Story (TEMAS) / WPS
TerraNova series / CTB
Test Anxiety Inventory (TAI) / CPP
Test of Cognitive Skills—Second Edition (TCS/2) / CTB
Test of Nonverbal Intelligence—Second Edition (TONI-2) / PRO-ED
Test of Variables of Attention (T.O.V.A) / UAD
Tests of Achievement and Proficiency / Riverside
Tests of Adult Basic Education (TABE) / CTB
The Holtzman Inkblot Technique (HIT) / TPC
Thematic Apperception Test (TAT) / Harvard
The Rosenzweig Picture-Frustration Study (P-F Study) / PAR
The Scenotest / H & H
The Values Scale / CPP
Vigil Continuous Performance Test (VIGIL) / ForThought
Vineland Adaptive Behavior Scales (VABS) / AGS
Washington University Sentence Completion Test (WUSCT) / Erlbaum
Wechsler Adult Intelligence Scale—Revised (WAIS-R) / TPC
Wechsler Individual Achievement Test (WIAT) / TPC
Wechsler Intelligence Scale for Children—Third Edition (WISC-III) / TPC
Wechsler Preschool and Primary Scale of Intelligence—Revised (WPPSI-R) / TPC
Wonderlic Personnel Test / Wonderlic
Woodcock-Johnson Psycho-Educational Battery—Revised (WJ-R) / Riverside
Woodcock Reading Mastery Tests—Revised / AGS
Work Keys / ACT

Apéndice B

Direcciones de editoriales, distribuidores y organizaciones relacionados con las pruebas

AAMR
- American Association on Mental Retardation
  444 North Capitol Street, N.W., Suite 846
  Washington, DC 20001-1512

ABPP
- American Board of Professional Psychology
  2100 East Broadway, Suite 313
  Columbia, MO 65201-6082

ACA
- American Counseling Association
  5999 Stevenson Avenue
  Alexandria, VA 22304-3300

ACT
- American College Testing Program
  ACT National Office
  2201 North Dodge Street
  P.O. Box 168
  Iowa City, IA 52243-0168

AERA
- American Educational Research Association
  1230 Seventeenth Street, N.W.
  Washington, DC 20036-3078

AGS
- American Guidance Service, Inc.
  4201 Woodland Road
  Circle Pines, MN 55014-1796

APA	American Psychological Association 750 First Street, N.E. Washington, DC 20002-4242
ASC	Assessment Systems Corporation 2233 University Avenue, Suite 200 St. Paul, MN 55114
ATP	Association of Test Publishers 655 Fifteenth Street, N.W., Suite 320 Washington, DC 20005
BoTA	Board on Testing and Assessment National Research Council 2101 Constitution Avenue, N.W. Washington, DC 20418
Buros	Buros Institute of Mental Measurements P.O. Box 880348 135 Bancroft Hall Lincoln, NE 68588-0348
CEEB	College Entrance Examination Board 45 Columbus Avenue New York, NY 10023-6992
CPP	Consulting Psychologists Press, Inc. 3803 East Bayshore Road P.O. Box 10096 Palo Alto, CA 94303
CPS	C.P.S., Inc. P.O. Box 83 Larchmont, NY 10538
CTB	CTB/McGraw-Hill 20 Ryan Ranch Road Monterey, CA 93940
EdITS	Educational and Industrial Testing Service P.O. Box 7234 San Diego, CA 92167
EEOC	Equal Employment Opportunity Commission 1801 L Street Washington, DC 20507

Erlbaum	• Lawrence Erlbaum Associates, Inc. 10 Industrial Avenue Mahwah, NJ 07430-2262
ETS	• Educational Testing Service Publications Order Services P.O. Box 6736 Princeton, NJ 08541-6736
	• ETS Test Collection Mailstop 30-B Rosedale Road Princeton, NJ 08541-0001
ForThought	• ForThought, Ltd. Nine Trafalgar Square Nashua, NH 03063
GRE	• Graduate Record Examinations Educational Testing Service P.O. Box 6000 Princeton, NJ 08541-6000
Harcourt Brace	• Harcourt Brace Educational Measurement Educational Testing Division of TPC 555 Academic Court San Antonio, TX 78204-2498
Harvard	• Harvard University Press 79 Garden Street Cambridge, MA 02138
HAS	• Hogan Assessment Systems, Inc. P.O. Box 521176 Tulsa, OK 74152
H & H	• Hogrefe & Huber Publishers United States Office: P.O. Box 2487 Kirkland, WA 98083 Swiss Office: Verlag Hans Huber Länggass-Strasse 76 CH-3000 Bern 9 Switzerland

IBM	•	IBM K-12 Education 4111 Northside Parkway Atlanta, GA 30327
Infantest	•	Infantest Corporation P.O. Box 18765 Cleveland Heights, OH 44118-0765
IPAT	•	Institute for Personality and Ability Testing, Inc. P.O. Box 1188 Champaign, IL 61824-1188
MRI	•	Management Research Institute, Inc. 6701 Democracy Blvd., Suite 300 Bethesda, MD 20817
NCME	•	National Council on Measurement in Education 1230 Seventeenth Street, N.W. Washington, DC 20036
NCS	•	National Computer Systems, Inc. P.O. Box 1416 Minneapolis, MN 55440
NFER-Nelson	•	NFER-Nelson Publishing Company, Ltd. Darville House, 2 Oxford Road East Windsor-Berkshire, SL4 1DF United Kingdom
Oxford	•	Oxford Psychologists Press, Ltd. Lambourne House 311 321 Banbury Road Oxford OX2 7JH England
PAR	•	Psychological Assessment Resources, Inc. P.O. Box 998 Odessa, FL 33556-0998
PDI	•	Personnel Decisions International 2000 Plaza VII 45 South Seventh Street Minneapolis, MN 55402-1608
PRO-ED	•	PRO-ED 8700 Shoal Creek Boulevard Austin, TX 78757-6897

- Riverside — The Riverside Publishing Company
  425 Spring Lake Drive
  Itasca, IL 60143

- RNL — Reitan Neuropsychological Laboratory
  2920 South Fourth Avenue
  Tucson, AZ 85713-4819

- *The Score* — Newsletter for Division 5 of the American Psychological Association
  4201 Woodland Road
  Circle Pines, MN 55014

- Sigma — Sigma Assessment Systems, Inc.

  United States Office:
  1110 Military Street
  P.O. Box 610984
  Port Huron, MI 48061-0984

  Canadian Office:
  Research Psychologists Press, Inc.
  650 Waterloo Street, Suite 100
  P.O. Box 3292, Station B
  London, ON N6A 4K3

- SilverPlatter — SilverPlatter Information
  100 River Ridge Drive
  Norwood, MA 02062-5043

- SIOP — Society for Industrial and Organizational Psychology, Inc.
  P.O. Box 87
  Bowling Green, OH 43402

- SPA — Society for Personality Assessment
  750 First Street, N.E.
  Washington, DC 20002-4242

- SRA — McGraw-Hill/London House
  SRA Business and Industry Assessments
  9701 West Higgins Road
  Rosemont, IL 60018-4720

- Stoelting — Stoelting Company
  620 Wheat Lane
  Wood Dale, IL 60191

Swets	Swets Test Services Heereweg 347b 2161 CA Lisse Nederland
Thomas	Charles C. Thomas Publisher 2600 South First Street Springfield, IL 62794-9265
TPC	The Psychological Corporation 555 Academic Court San Antonio, TX 78204-2498
UAD	Universal Attention Disorders, Inc. 4281 Katella #215 Los Alamitos, CA 90720
UIP	University of Illinois Press 1325 South Oak Street Champaign, IL 61820
UMP	University of Minnesota Press Test Division 111 Third Avenue South, Suite 290 Minneapolis, MN 55401
USES	United States Employment Service Western Assessment Research and Development Center 140 East 300 South Salt Lake City, UT 84111
U.S. Military	United States Military Entrance Processing Command Attn.: Technical Directorate 2500 Green Bay Road North Chicago, IL 60064-3094
Wonderlic	Wonderlic Personnel Test, Inc. 1509 North Milwaukee Avenue Libertyville, IL 60048-1380
WPS	Western Psychological Services 12031 Wilshire Boulevard Los Angeles, CA 90025-1251

Índice de instrumentos

Batería de Aptitud Vocacional de las Fuerzas Armadas (ASVAB), 193, 271, 278, 500-502, 553
Batería de Aptitudes Multidimensionales (MAB), 290-293, 318, 555
Batería de Clasificación de la Marina, 18
Batería de Evaluación para Niños de Kaufman (K-ABC), 222-224, 236, 555
Batería de Pruebas de Aptitudes Generales (GATB), 26, 159, 499-500, 527, 554
Batería de Pruebas Neuropsicológicas de Halstead-Reitan (HRB), 519, 554
Batería Neuropsicológica Luria-Nebraska (LNNB), 519, 555
Batería Psicoeducativa de Woodcock-Johnson, Revisada, 239n, 522,
Baterías de aprovechamiento general, 481-483
Baterías de niveles múltiples, 278-286

Cuestionario de Adaptación de los Estudiantes a la Universidad (SACQ), 534, 556
Cuestionario de Dieciséis Factores de Personalidad (16PF), 363-364, 556
Cuestionario de Estilos de los Estudiantes, 449, 556

Escala de Binet-Simon, 37-38, 55, 205
Escala de Autoconcepto del Estudiante (SSCS), 456-457, 556
Escala de Autoestima (SES), 456
Escala de Conceptos Básicos de Bracken, 490, 554
Escala de Ejecución de Pintner-Paterson, 254
Escala de Ejecución para Niños Sordos de Wechsler, 254-255
Escala de Habilidad Diferencial, 231-232
Escala de Inteligencia para Adultos de Wechsler, Revisada (WAIS-R), 75, 215-216, 290-291
Escala de Inteligencia Stanford-Binet, 38, 40, 42, 236, 513, 556
Escala de Inteligencia Táctil, 257n, 554
Escala de Inteligencia Wechsler-Bellevue, 215
Escala de Madurez Mental de Columbia (CMMS), 259, 554
Escala de Madurez Social de Vineland, 249
Escala de Valores, 387, 556
Escala Ejecutiva de Arthur, 254
Escala Guttman, 57n, 77n
Escala I-E (Escala Interno-Externo), 408-409
Escala Internacional de Ejecución de Leiter, Revisada, 261-263, 555
Escala Internacional de Ejecución de Leiter, 258
Escala Nacional de Referencia, 70
Escala tipo Likert, 406-407
Escala tipo Thurstone, 406
Escalas absolutas, 174-177
Escalas Británicas de Habilidad (BAS), 226, 554
Escalas de Bayley, sobre el Desarrollo Infantil, 237-239, 551
Escalas de Binet, 37-38
Escalas de Clima Social, 462-463, 556
Escalas de Conducta Adaptativa (ABS), 251-252, 553
Escalas de Conducta Adaptativa de Vineland (VABS), 249-251, 556
Escalas de desarrollo de Kaufman, 222
Escalas de Habilidad Diferencial (DAS), 42, 192, 236, 318, 554
Escalas de Habilidades Infantiles de McCarthy (MSCA), 239-240, 555
Escalas de Inteligencia, 62, 65, 512
Escalas de Inteligencia de Wechsler, 62, 65, 513
Escalas de Inteligencia para Edades Preescolar y Primaria de Wechsler (WPPSI, WPPSI-R), 216, 236, 556
Escalas de Inteligencia para Niños de Wechsler (WISC-R, WISC-III), 75, 215, 216, 556
Escalas del Desarrollo Psicológico Infantil, 241-242, 554
Escalas ordinales, 56-58

Escalas Ordinales del Desarrollo Psicológico, 241-242, 555
Escalas ordinales piagetianas, 127
Escalas piagetianas, 240-244
Escalas R-I-A-S-E-C, 403-404

Inventario Básico de Personalidad (BPI), 358, 373-374, 553
Inventario Básico de Personalidad de Jackson, 358, 553
Inventario Breve de Síntomas (BSI), 350, 554
Inventario Clínico para Adolescentes de Millon (MACI), 369-370, 555
Inventario de Adaptación Transcultural (CCAI), 534, 554
Inventario de Ansiedad (TAI), 22, 383, 556
Inventario de Autoconcepto Personal y Académico (PASCI), 457
Inventario de autodescripción, 44.
Inventario de Consumo del Alcohol, 525, 553
Inventario de Creencias Profesionales, 528, 554
Inventario de Depresión de Beck (BDI), 525, 553
Inventario de Desarrollo Profesional, 527, 554
Inventario de Empleo PDI, 510, 555
Inventario de Estados y rasgos de Ansiedad (STAI), 384, 556
Inventario de Estados y rasgos de Ansiedad para Niños (STAIC), 384, 556
Inventario de Estados y rasgos de Expresión de la Ira (STAXI), 384, 556
Inventario de Evaluación de Carreras, versión ampliada (CAI-EV), 400n, 554
Inventario de Evaluación de Carreras, versión vocacional (CAI-VV), 400, 554
Inventario de Evaluación de la Personalidad (PAI), 358-359, 555
Inventario de Intereses de Strong (SII), 389-396, 556
Inventario de Intereses Profesionales (CII), 396, 527, 554
Inventario de la Calidad de Vida (QOLI), 533, 555
Inventario de Observación en el Hogar para la Medición del Ambiente (HOME), 270
Inventario de Orientación Vocacional, 403, 554
Inventario de Personalidad de Hogan (HPI), 508, 555
Inventario de Personalidad NEO, Revisado (NEO PI-R), 365-366, 555
Inventario de Personalidad para Adolescentes de Millon (MACI), 369-370, 555
Inventario de Personalidad para Jóvenes (PIY), 362, 468n, 555
Inventario de Personalidad para Niños, 360-362
Inventario de Personalidad para Niños, (PIC y PIC-R), 361-362, 468n, 555
Inventario de Personalidad de Jackson, Revisado (JPI-R), 373, 554
Inventario de Preferencias Personales de Edwards (EPPS), 370-371, 377, 554
Inventario de Servicios al Cliente PDI, 510, 554
Inventario de Valores para la Vida, 386-387
Inventario Multiaxial Clínico -III de Millon (MCMI-III), 367-370, 555
Inventario Multiaxial Clínico de Millon, 367-370
Inventario Multifásico de Personalidad de Minnesota (MMPI, MMPI-2, MMPI-A), 74, 351-359, 555
Inventario para la Personalidad Inadaptada y Adaptada (SNAP), 369n, 555
Inventario Psicológico de California (CPI), 359-360, 554
Inventarios autodescriptivos, 44, 362-367
Inventarios de autoestima, 455-457
Inventarios de personalidad, 362-374

Lista de Verificación de 90 Síntomas, Revisada (SCL-90-R), 349-350, 556

Prueba de asociación de palabras, 425-426
Prueba de Asociación Libre de Kent-Rosanoff, 426
Prueba de Asociación Libre, 44, 425-426
Prueba de Dibujo de Goodenough-Harris, 264-266, 430, 554
Prueba de Dibujo de un Hombre de Goodenough, 19, 264-266
Prueba de Dibujo de un Hombre, 19, 264-266, 554
Prueba de Dibujo de una Persona de Machover (D-A-P), 429-430, 438, 555
Prueba de la cesta, 493
Prueba de la Estructura de las Habilidades de aprendizaje del Intelecto, 314, 556
Pruebas Alfa y Beta del Ejército Estadounidense, 39, 234, 271
Pruebas comprensivas de habilidades básicas, 192, 554
Pruebas de aptitudes especiales, 39-42, 502-505
Pruebas de aptitudes mecánicas, 504

Pruebas de Calificación de las Fuerzas Armadas (AFQT), 163, 271, 500, 553
Pruebas de aprovechamiento, 475-478
Pruebas de competencia mínima en habilidades básicas, 483-484
Pruebas de Evaluación Académica (SAT), 62, 556
Pruebas de Integridad, 508
Pruebas de Lápiz y Papel, 503-504, 507
Pruebas de Maduración, 489-490
Pruebas de Poder, 102
Pruebas de referencia a criterio, 76

Técnica de elección forzada, 376-377
Técnica de Manchas de Tinta de Holtzman (HIT), 418-419, 556
Técnica de representación de papeles (Roleplaying), 453-454
Técnica nominativa, 467-468
Técnicas de desempeño de juego y juguete, 431-432
Técnicas de Dibujo, 429-431
Técnicas de ejecución, 429-432
Técnicas de evaluación de los problemas de aprendizaje, 519-521, 522
Técnicas de evaluación, 521-524
Técnicas de Incidentes Críticos, 493
Técnicas de la Entrevista, 464-465
Técnicas de Manchas de Tinta de Holtzman (HIT), 418-419, 556
Técnicas de Manchas de Tinta de Rorschach, 5, 195, 411-414
Técnicas pictóricas, 419-425
Técnicas verbales, 425-427
Test Adaptado a las Computadoras (TAC), 192, 276-278
Test Autoaplicado de Habilidad Mental de Otis, 272, 498
Test Breve de Inteligencia de Kaufman (K-BIT), 225, 555
Test de Apercepción Gerontológica, 423
Test de Apercepción Infantil (CAT), 422, 522
Test de Apercepción para Ancianos, 423, 556
Test de Apercepción para Niños de Roberts (RATC), 422, 555
Test de Apercepción Temática (TAT), 419-423, 556
Test de Aprovechamiento Educativo de Kaufman (K-TEA), 521, 555
Test de Aprovechamiento Individual de Wechsler (WIAT), 521, 556
Test de Aptitud Diferencial (DAT), 288-290, 554
Test de Aptitud para el Aprendizaje de Hiskey-Nebraska, 255
Test de Aptitud para el Aprendizaje de los Ciegos (BLAT), 256-257
Test de Clasificación Pictórica, 259
Test de Completamiento de Oraciones de la Universidad de Washington (WUSCT), 427n, 454-455
Test de Comprensión Mecánica de Bennett, 505-506, 553
Test de Conceptos Básicos de Boehm, 490, 554
Test de Desarrollo Educativo de Iowa, 482, 555
Test de Desempeño en Vigilia Continua, 532, 556
Test de Destreza con Piezas Pequeñas de Crawford, 503, 554
Test de Destreza Motora de Bruininks-Oseretsky, 252, 554
Test de Destreza Motora de Oseretsky, 252
Test de Diagnóstico de Lectura y Matemáticas de California, 488, 554
Test de Educación Básica para Adultos (TABE), 484, 556
Test de Evaluación Escolar (SAT), 62, 556
Test de Figuras Encajadas, 446-447, 554
Test de Habilidad de Aprendizaje Espacial, 201
Test de Habilidad Escolar de Otis-Lennon, (OLSAT), 280, 281, 286, 482, 555
Test de Habilidades Básicas de Iowa, 482, 555
Test de Habilidades Cognoscitivas (CogAT), 280, 281, 284, 286, 554
Test de Habilidades Cognoscitivas (TSC), 280, 283, 285, 286, 556
Test de Habilidades Mentales para los Adultos de Schaie-Thurstone, 334n, 556
Test de Inteligencia Culturalmente Justo de Cattell, 311, 554
Test de Inteligencia Infantil de Fagan, 245, 554
Test de Inteligencia No Verbal (TONI-2), 266
Test de Inteligencia para Adolescentes y Adultos de Kaufman (KAIT), 224-225, 555
Test de Inteligencia para Ciegos de Perkins-Binet, 256
Test de Inteligencia para Niños con Deficiencias Visuales, 257
Test de la Estructura de las Habilidades de Aprendizaje del Intelecto, 314, 556
Test de Pasos Seriados de Adición Autiditiva (PASAT), 532, 555
Test de Personal de Wonderlic, 498, 556
Test de personalidad, 44-45
Test de Pronóstico de Álgebra de Orleans-Hanna, 489, 555
Test de Repertorio de Construcción de Roles, 460-461

Test de Retención Visual de Benton (BVRT), 518, 553
Test de Variables de Atención (TOVA), 531, 556
Test de Vocabulario con Ilustraciones de Peabody (PPVT-R), 251, 258-259, 555
Test del Grupo de Figuras Encajadas, 446, 447, 448, 554
Test Diagnóstico en la Lectura de Stanford, 488, 556
Test Diagnóstico en Matemáticas de Stanford, 488, 556
Test Gestáltico de Bender (BGT), 518
Test Gestáltico Visomotor de Bender, 518, 553
Test KSA para el trabajo en equipo, 507, 556
Test para Empleados de Oficina de Minnesota (MCT), 506, 555
Tests de Apercepción Temática, 419-423
Tests de Aprovechamiento de California, 192, 482, 554
Tests de Habilidad de Lectura de Woodcock, Revisados, 232, 556
Tests referidos a dominio, 76-80

PRENTICE
HALL